CRC PRESS ▪ TAYLOR & FRANCIS

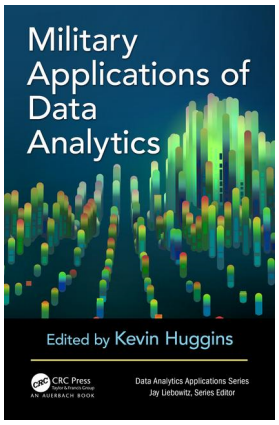
Data Analytics Applications

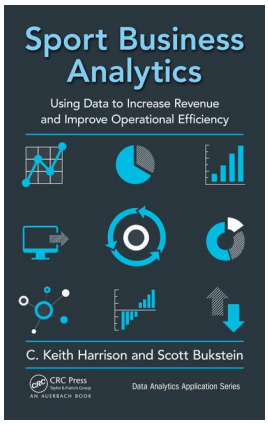
A Chapter Sampler

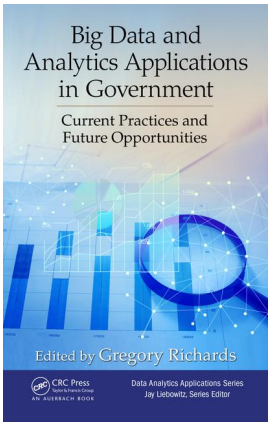
CRC Press
Taylor & Francis Group

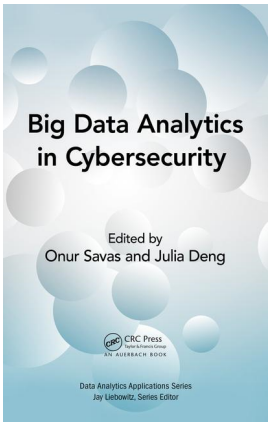
www.crcpress.com

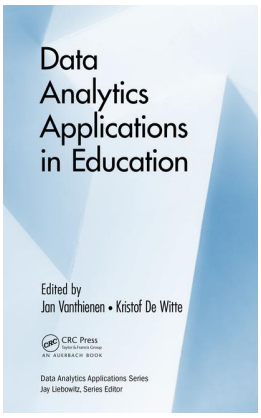
Contents

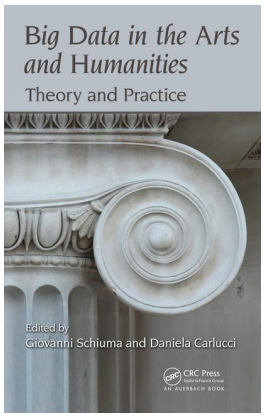












Chapter 2

Network Modeling and Analysis of Data and Relationships: Developing Cyber and Complexity Science

Chris Arney, Natalie Vanatta, and Matthew Sobiesk

Contents

Introduction

Networks are not only ubiquitous, but also lie at the core of the economic, political, military, and social fabric of modern society. As stated in a National Research Council (2005) report, "society depends on a diversity of complex networks for its

very existence." Today's scientists and analysts perform network modeling and data analytics in many applications. Network science (NS) as a component and partner of data analytics (DA) is critical to understanding many military-relevant issues, such as (Brandes et al., 2013):

- Communication flow
- Command and control
- Managing unit operations
- Implementing information assurance
- Modeling terror cells and their processes
- Gathering and processing intelligence
- Maintaining security in physical and informational systems
- Enabling engagement within the Army's Global Landpower Network
- Decision making

In health and biology, network applications include mapping genetic and protein networks and their roles in disease, representing the brain morphology by networks, forecasting and diagnosing disease contagion, and analyzing various levels and regions of ecosystems. Network-based applications involving physical networks include managing communication and computer systems, operating logistics and transportation systems, and designing infrastructures in various buildings, structures, and systems (e.g., water, waste, and heat). The network models associated with social processes involve conducting collaborative decisions and group learning, modeling collective/team behaviors that involve coordinated activities, modeling the meanings of textual and spoken language to understand influence, and tracking the emergence of societal impact and achievement.

Network models can merge the social, informational, communication, and physical layers of a system or organization into an interconnected, unified system for a broad, integrated analysis (Arney, 2016; Arney and Coronges, 2015). The network layers collectively produce an all-encompassing, non-reductive model that permits multiple scalings, processing of tremendous volumes of data, suitable system complexity, and appropriate diversity and specializations within the system or organization. Network modeling enables the dynamics in the structures and processes of the phenomenon being modeled to build usable knowledge and results for viable decision making. One of the recent results in NS shows the impossibility of perfect control over organizational and entity behavior in social networks (West, 2015). A subtle hand of delicate management through shared vision and autonomy is often more powerful than rigid micro-control through rules, regulations, and detailed instructions. This result is significant for military leader networks. As Roehner (2007) reflected about these data-driven contexts: "the real challenge is to do real physics and real sociology in the framework of network theory."

Data analytics (DA) encompasses even more techniques and methods than NS. The National Academies (2017, p. 64) describes and categorizes DA as:

- Descriptive and exploratory: "Using data to summarize and visualize the current state."
- Predictive: "Using data to determine patterns and predict future outcomes and trends."
- Prescriptive: "Using data to determine a set of decisions … that give rise to the best possible results."

The techniques used to perform these DA components are data extraction, natural language processing, data mining, machine learning, statistical analysis, stochastic modeling, regression, optimization, simulation, clustering, and classification. The combination of both unstructured text and highly structured quantitative data sometimes is referred to hard-soft data fusion.

We will report on network modeling and data analytics as they relate to two basic-science issues with associated applications within the military—cyber science and complexity science. In particular, we will outline the type of DA and NS work being performed and provide examples in both areas. Cyber science often relies on descriptive and predictive data analytics to find anomalies in active network processing and prescriptive data analytics to determine actions needed to protect or fix attacked networks. The U.S. Department of Defense (DoD) Cyber Strategy (2015) guides much of the efforts we have assembled in our modeling and problem solving of cyber and information science problems.

Cyber Science

One area where we use NS and DA is in cyber problem solving. We use network modeling to build a framework for cyber problems using game theory concepts linked to mathematical topology (information networks) and cultural modeling (social networks).

In 2006, the Joint Chiefs of Staff decreed the Fifth Operational Warfighting Domain for the United States military: the cyber domain. At times, a seemingly made-up word, cyber is still not clearly defined and understood a decade later. Academia, industry, popular media, and government can all debate the correct usage of the noun, but for purposes of this discussion, we will use definitions from the military community. *Joint Publication 3-12 (R) Cyberspace Operations* of the U.S. DoD (2013) defines cyberspace as "the global domain within the information environment consisting of the interdependent network of information technology infrastructures and resident data, including the Internet, telecommunications networks, computer systems, and embedded processors and controllers." The Joint Publication further comments that cyberspace has multiple layers: "physical network, logical network, and cyber-persona" layers. The physical network layer consists of the geographic

components that produce the medium where the data travel. The logical network layer is an abstraction of the information space (e.g., the URL for a website, no matter where it resides in the physical network). The cyber-persona layer represents an even higher level of abstraction of the logical network (e.g., the website itself and the people that build and operate the network). As the military continues to explore the cyber domain (both with an offensive and defensive mindset), the principles of network science and the analysis of data play a large role in their understanding.

In general, cyber operations are not constrained by geographic terrain but are bounded by the network connections and the cultural space of the operational mission and forces. Often the topology of the network dramatically affects the operational plan and situational awareness. Primary concerns in cyber situational awareness are defining information needs and the types of activities involved in the operation. The topology of the network models for information operations is much like a terrain map or operational overlay for physical operations. Understanding the cultural domain involves analysis of human and technical factors, such as computer and network platforms, cultural bias, ability to collect relevant and timely data, and legal issues. On the analytic side, the technical work involves analyzing the data to show the connections of structures and processes, and identifying possible patterns in the dataset. The methodological tools in network modeling including machine learning, qualitative and quantitative analysis, and statistical evaluation (Carter et al., 2014; Mayhew et al., 2015).

Today's military cyber forces have three primary missions: (1) defend DoD networks, systems, and information, (2) defend the U.S. homeland and U.S. national interests against cyberattacks of significant consequence, and (3) provide cyber support to military operational and contingency plans. Under these missions, cyber forces must be able to understand the current network infrastructure that is used both for command and control capabilities but also for watching the latest cat videos on YouTube. See Figure 2.1 for a framework for such a network structure.

One interesting challenge in understanding the infrastructure is the development of a PACE (primary, alternate, contingency, and emergency) plan for communications. These should be four separate means of organic communication capabilities that units can fall back through in case of transmission disruption. Traditionally, this might be secure FM radio, then high-frequency radio, then secure satellite phone, and finally cellular phone. With the many layers of abstraction in our evolving communications technology, it is even difficult for experts to understand the potential points of failure and interdependencies within transmission spectrums in these extremely complex environments. For example, there is a security operations center supporting multiple states that recently fell prey to this complexity. They built the center with extreme redundancies: dual ISPs providing bandwidth over two different fiber lines that entered the facility from completely different cardinal directions. They did everything to ensure separate communications paths to protect their operations, and yet it turned out that within 50 miles from their facility, both links went over a single bridge. NS modeling can be used to visualize and simulate the complex communications spectrum to ensure that

Figure 2.1 Interaction of cyber domain aspects.

military PACE plans are accurate and actionable. Examination of multimodal transportation systems uses a set of subnetworks (each representing a transportation method such as plane, train, car, boat, bicycle, and walking) and then builds a logical network to describe connections between the layers. These same techniques could be used to describe the layers of communication spectrums (such as cellular, high frequency [HF], ultrahigh frequency [UHF], wireless, satellite) and resulting physical infrastructure. Then, apply tools of percolation theory and cascading failures to determine viability of the PACE plan during operations.

Additional NS and DA skills can be focused on the data storage and processing problem. On January 10, 2017, Army Secretary Eric Fanning directed that 60% of the Army's 1200 data centers must be closed by the end of 2018 and 75% by 2025. This is an effort to consolidate services to improve security and workforce efficiencies. This translates into a network optimization problem spanning the globe. Where do you place certain services to cover the military installations across continents? A current data center could be providing 1–2 services or up to 40+ different services to individuals within a mile of the location or spanning continents. With different characteristics of services, how to design an optimized location of the remaining data centers that can still meet the operational needs for the military both when they are working in garrison and when they are deployed in the field is an interesting problem.

Given the complexity and ad hoc nature of most military network infrastructure, it becomes difficult to model and/or simulate the normal behavior of traffic (Paxton

et al., 2014). Lack of robust modeling and DA capabilities hinders leaders' abilities to make strategic risk decisions, quantify the impact of degraded systems, and identify cyber key terrain. When troubleshooting communications networks, often the issue and problems that need to be addressed in the operational environment are hidden and only the symptoms are visible. The determination of the cause of the problem—a bug, an innocent human error, bad hardware, multi-layered firewall, a dynamic, fast-acting AI, or a malicious attack—is often undetectable without clever and innovative use of DA. DA can also provide insights into the optimal placement of sensors within networks to detect maliciousness without undermining efficient routing of traffic. NS techniques can also help reverse engineer the infection of networks by malware. Cyber analysis is needed for time-sensitive problem solutions and dynamic network performance that are not as common in other domains.

Complicating the cyber forces' defensive mission is the massive amount of data that is generated by networks, systems, and people. The search for the needle in the haystack (solution) is difficult due to the volume of noise. DA techniques coupled with behavioral understanding are required to turn data into information and then into actionable intelligence. A defender must look not only for the outside threat but also for the insider threat. Network traffic sensors (aimed at detecting abnormal patterns), physical security logs, system access logs, and many more data streams (both technical and social) are fed into "Big Data" platforms. Then, using machine learning techniques, a system can detect insider threat instances from a non-human perspective with surprising accuracy, drawing connections that were not previously considered relevant.

The intersection of large datasets and machine learning is causing an evolution of data analytics. A large information-based company processes 600 billion events a day (approximately 3 petabytes of data a day) through a single pipeline and then uses an assortment of machine learning techniques to sort events, whether they are processed in real time or batch processed at a later time. The system learns situational behavior to make autonomous decisions at any given moment in any given situation about what data should be immediately processed to provide the best real-time quality of service to customers. While the military insider threat problem described in the previous paragraph would result in significantly less than 600 billion events a day for a given network region, the ability to sort data at machine speed for various processes based on the developing situation would provide quicker indications of malicious actors on the network.

NS concepts also provide cyber defenders understanding of the network attack surface. Attack graphs as shown in Figure 2.2 represent the paths through a network that end in a state where the intruder successfully gains access to protected systems (Ingols et al., 2009). The size and complexity of modern networks make it difficult to manually determine/maintain an account of the exploitable surfaces. Researchers are exploring new techniques and algorithms to overcome the complexity of the state explosion problem, which occurs when the number of vulnerabilities in a network grows large. Coupling attack graphs with machine learning capabilities allows for models to also hypothesize on missed vulnerable paths and/or missed alerts that intrusion detection systems don't see.

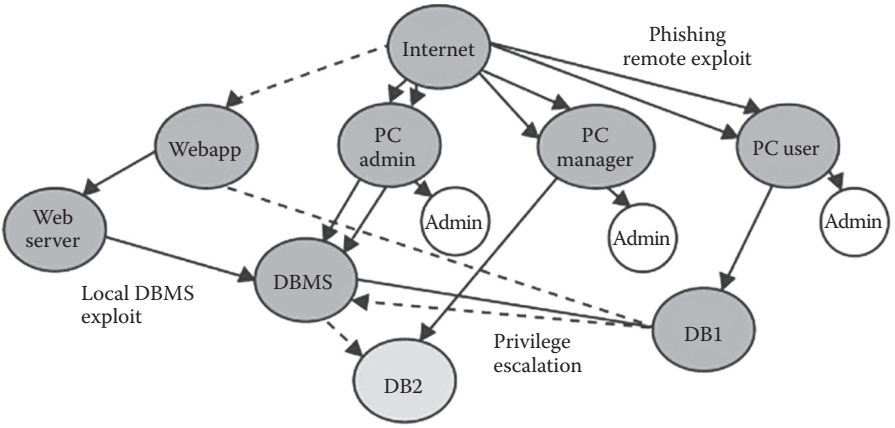


Figure 2.2 Network attack graph.

One element of the complexity of cyberspace is identifying the problem to be solved or the issue to be confronted. Another element is the underlying competitive nature of the attack-defender dynamic. Hackers and malicious systems are pitted against defenders of information and systems performance. Attacking elements often try to hide their true identity. Defenders try to hide or change their patterns so attackers cannot find weaknesses. This game theoretic setting takes modeling to new heights in what-if, cause-effect, and forensics questions. Cyber problem solving is an unstructured process that often requires high-dimensional data, nonlinear models, and dynamic modification to adapt to the constantly changing situations.

An analogy of cyber systems to biological systems can be helpful. Monocultures are efficient but vulnerable because uniformity and patterns create a form of weakness to non-normal conditions. Polycultures and diversity are inefficient, but usually robust to a changing environment and therefore survivable (Shah, 2014). The result is that diversity creates a form of strength. The ultra-efficiency of uniform order can produce fragility. This idea of adding randomness and diversity is not intuitive to modelers who have worked in other domains. Fortunately, despite our intent to make the Internet more uniform and efficient, it is not. The Internet works because of its inherent diversity and randomness. Yet we continue to follow our intuition and design our networks and systems primarily for efficiency. The result can be super-efficient networks that are often rigid and brittle. Even today, we revert to old, yet unhelpful, habits. When things go wrong in our cyber systems, we react by enforcing rigid discipline and control that destroys diversity and ultimately hurts the robustness and resilience. So, when weaknesses are found, cyber scientists may need to design in more intentional randomness into the network. Modelers use the system's diversity for improved survivability at the cost of efficiency. Randomness means that no one (not even the designer or builder) has precise control, but overall performance will still be higher than highly patterned, overprogrammed, inflexible, and eventually broken systems.

What makes a network robust, survivable, and hard to kill paradoxically also makes it inefficient, difficult to manage, and vulnerable to penetration. Evolutionary biology shows that inherent diversity provides reliability at a price of some inefficiency. Evolutionary biology also teaches that change (adaptation and randomness) is needed in order to survive. Today's cyber systems are vulnerable and unpredictable—a place where actions and events happen fast. So, to survive on the network, you have to be able to react quickly and effectively—sometimes proactively, sometimes reactively. Diversity is the model attribute that best provides the potential for resilience to vulnerabilities and yet maintains the agility to change fast. One natural way to create diversity in cyber systems is through randomness (explicitly designed random processes). Nature provides diversity in its DNA and cells; cyber scientists need to build diversity and randomness into their systems.

Cyber modeling enhanced by computational game theory and simulation enables war gaming of the basic elements of the cyber competition. These games and simulations are used to test capabilities, probe for vulnerabilities, fix performance degradation, and exercise the cyber systems. These are the research tools needed to enhance cyber security. Artificial intelligence techniques like machine learning and reinforcement learning are also elements in the models of the cyber framework. These faster-paced simulations will test the more advanced techniques of attacking and defending.

At the intersection of game theory and cyber security, models for static physical security games, such as Stackelberg Security Games, where defenders set a strategy and attackers surveil and attack, can be modified with more active defenders (Delle Fave et al., 2015; Shieh et al., 2016; Sinha et al., 2015). This framework gives a start for designing a fluid cyber situation with a defender trying to protect weighted targets, such as data servers, high-valued communication nodes, and physical access links. Attackers use probes, malware, exfiltration, and spoofing to attack the network. Defenders use honey pots, auditing, access control, detectors, and randomization of allocations and resources. These fluid actions are simulated so measures of the network and the strategies can be monitored and modeled for forensic analysis.

Researchers also experiment with smart cyber information systems with various security systems and test the use of randomness on performance and security measures. These tests can validate the framework and value of randomness in security. The goal of this computational approach is to develop measures of efficiency. The algorithms will defend or attack networked systems, predict and defend against attacks, respond to problems, use flexible and random designs, protect information, and monitor performance.

This information, coupled with machine learning techniques, could create autonomous defensive positions within cyberspace that, as attackers' techniques change, could modify a network's defenses without the intervention of a human. Being able to operate and respond at machine speed to threats is what military cyber forces require and what NS and DA can provide.

Ultimately, cyber science as a form of information science is different from the traditional sciences in terms of its competitive nature and highly structured network. Network models, Big Data analysis, game theory, and artificial intelligence are applied components of the digital cyber world. Computational game theory on a network is an important component of understanding the essence and dynamics of cyber science. As in many competitive security applications, the avoidance of operational and structural patterns through the inclusion of randomness and diversity are fundamental elements of cyber science. Therefore, this makes complexity an inherent element in cyber problems and their analysis. Cyberspace itself is the combination of many digital-related components that store, process, secure, protect, transmit, and use information. Network models establish a framework that incorporates the technical aspects of cyber operations, along with many human-based disciplines such as philosophy, ethics, law, psychology, policy, and economics that contribute to cyber analytics. Cyber science offers concepts and tools to understand complex security issues. In every domain and subject, there are differences in how data analytics and network modeling are used. Network models represent the connected elements and capture the dynamic of the attacker–defender interface. The cyber components that we study through NS and DA include:

- Authentication procedures
- Connections
- Operating systems
- Protocols
- Topology

There is an evolving future for data analytics and network modeling in cyber science. Cyberspace does require new, original ways of thinking and building models for the tasks that are part of this rapidly changing science. It is not often that the way ahead in a science is to implement randomness, embrace diversity, accept inefficiency, tolerate complexity, and thrive through interdisciplinary study. Cyber science in this form will be a challenge for our analytic community to accept, understand, and develop.

In a recent cyber-related application, a student at the United States Military Academy is using NS principles to explore whether submarine communications cables could pose a national security threat. Ninety-seven percent of the world's telecommunications traffic traverses these cables (shown in Figure 2.3), which are privately owned (Burnett et al., 2014). By creating a multilayer graph, he is exploring the interactions between the physical locations of the cable terminations, the multinational ownership conglomerations of the cables, and alliances between nation-states. History is full of examples of nation-states that tap these cables to steal information or deny access. NS will facilitate the simulation of a country denying access (removal of nodes and edges) and the second-/third-order effects to the ability to pass military operational traffic.

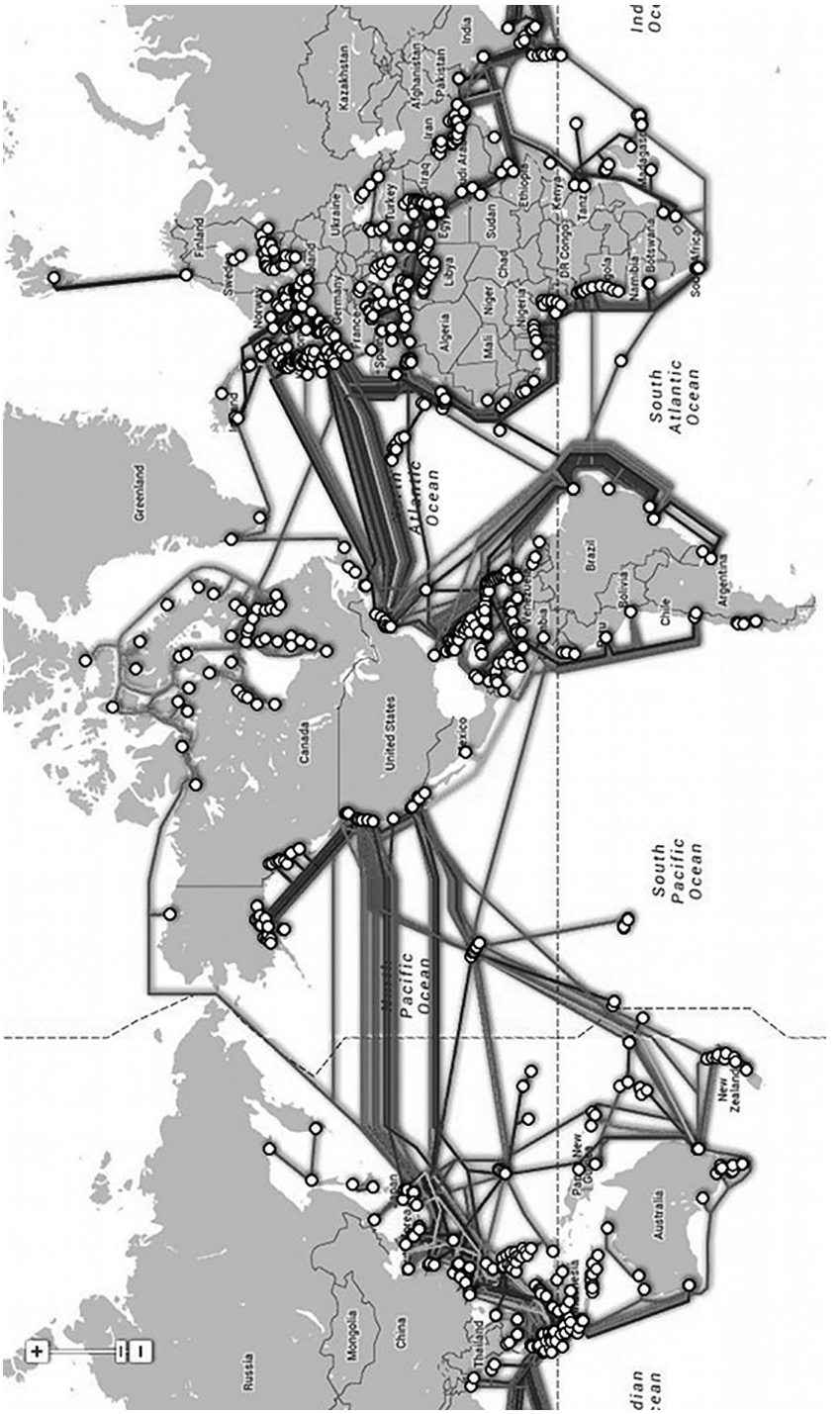


Figure 2.3 Snapshot of submarine communications cables. (From www.submarinecablemap.com.)

In another cyber-related application, analysts collect and study Twitter data to analyze the onset and evolution of social movements (Brantly, 2017; Korolov et al., 2016). They developed a framework to identify protest mobilization in social media and assess the likelihood of the protest occurring for a specified cause at given geographic location. Using machine learning, natural language processing, sentiment analysis, and influence network metrics, they study the tipping point and characteristic sentiment in the cyber domain to a kinetic action. By considering attempts to incite the protest or to influence public opinion via social media as a cyber-attack, they design methods to identify factors and mechanisms of protest development to suggest interventions. Future work will focus on identification of the antecedents of protest.

Complexity Science

The pace of discovery and progress in mathematics, science, society, and education continues to advance as the information, techniques, issues, and important questions shift (West, 2015). Science may be an evolutionary process, but the ways humans perform it, utilize it, and understand it can be revolutionary (deSolla Price, 1986; Weaver, 1948). Elements of this shift include: moving from little science (single investigators) to big science (large institutions supervising and controlling large groups) to team science (multidisciplinary, multi-organization, multiskilled, multination all-star teams of scientists); the increased role of complexity; connections in understanding the informational and networked nature of science; and the human utilization of science within the context of society (Holtz, 2015; Ledford, 2015; Scientific American Editors, 2015). Modern science through NS and DA is building a collective power that is more creative, more original, and more effective than the single disciplinary perspectives of the past. Some of this shift is attributed to the development of fractals and fractional mathematics (fractional calculus with fractional networking and fractional statistics) playing fundamental roles. Scientists hope to contribute to science's system of shared knowledge, appropriate levels of abstraction, and an enlightened human context that enables science to engage in the most compelling issues of society. It is no accident that this movement toward the integration of science coincides with a globalization of efforts and the proliferation of networks.

In our modeling and analysis efforts, we build models to understand the empirical differences in the shift of science. For example, we compared the collaboration networks for two exemplars of the shift (Paul Erdős [circa 1935–1995] and László Barabási [circa 1995–2015]). The networks we analyzed were the coauthor networks of the primary with the primary excluded. These are networks of co authorships of the coauthors, or the two degrees of separation network of the primary's coauthors. The differences in these networks indicate significant and dramatic change is taking place in scientific research, demonstrating the changing methods of science.

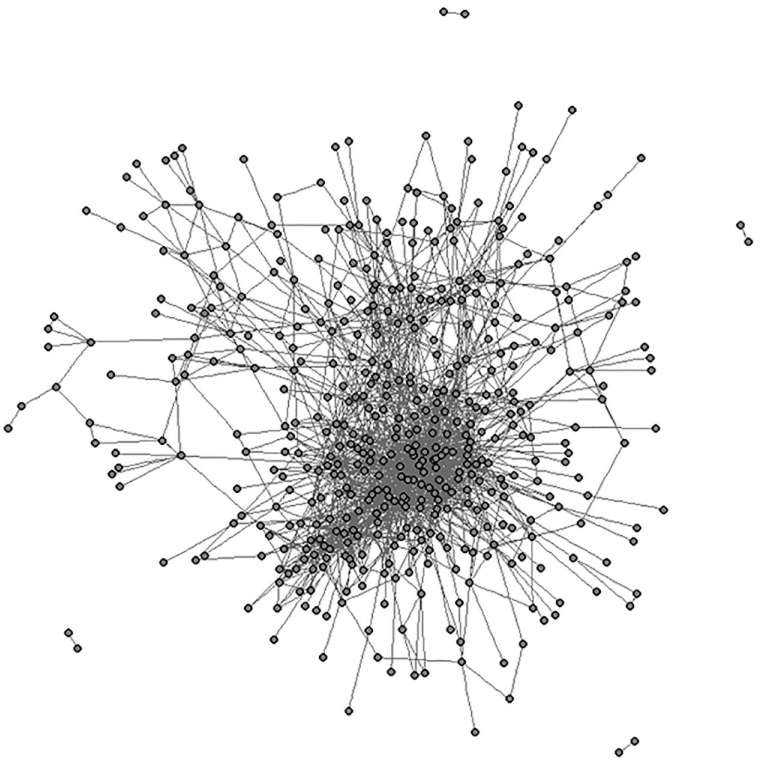


Figure 2.4 Erdős collaboration network.

For example, the Erdős network (a small science network) shown in Figure 2.4 has a much different structure core and peripheral structure than the Barabási network (team science), as shown in Figure 2.5. The Erdős network is also much less dense because of the smaller groups (often one person) that Erdős worked with.

The degree distribution graphs also show these same fundamental differences. The Erdős collaboration network distribution is shown in Figure 2.6 and the Barabási network distribution in Figure 2.7. The Barabási network distribution contains many more large-degree nodes enabling and demonstrating the team science approach to his research and problem-solving efforts.

The data summary for the two networks is provided in Table 2.1. The Erdős co authors have a mean degree of 6.76, whereas the Barabási co authors are much more connected with a mean degree of 28.4 and much more clustered into dense subnetworks with a much higher cluster coefficient.

DA and NS in the context of team science help us to understand the next layers of modern science that are the result of complexity (often created by unseen networks

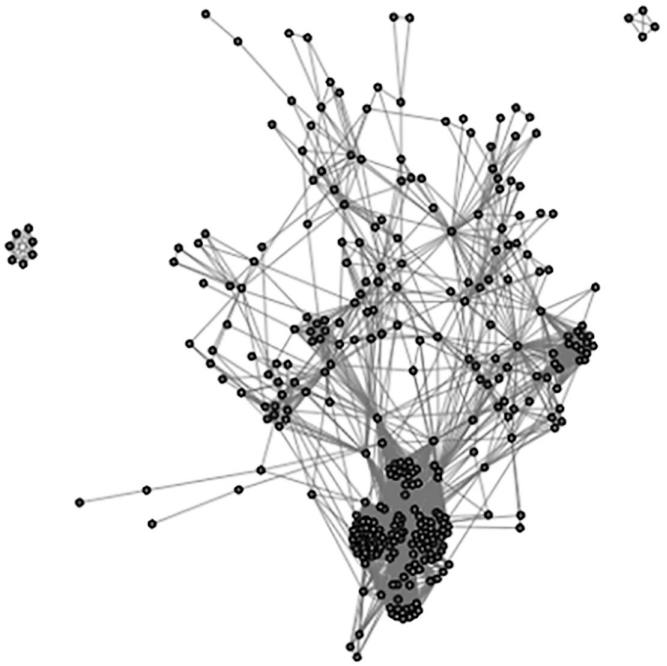


Figure 2.5 Barabási collaboration network.

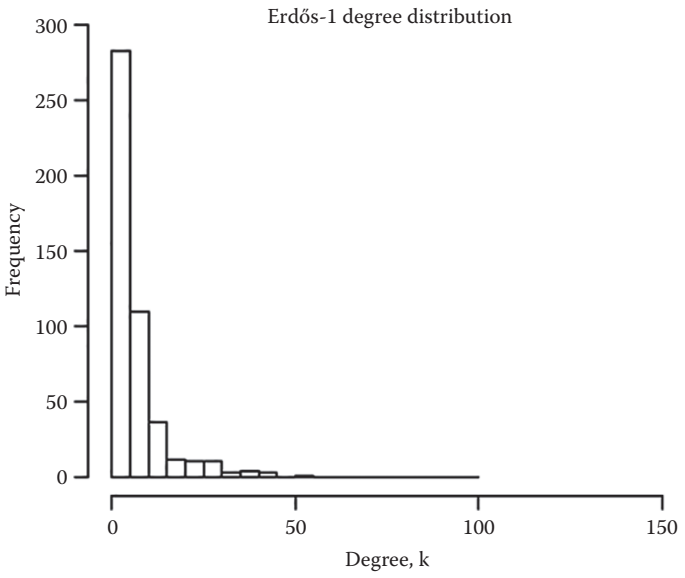


Figure 2.6 Distribution of nodes in the Erdős collaboration network.

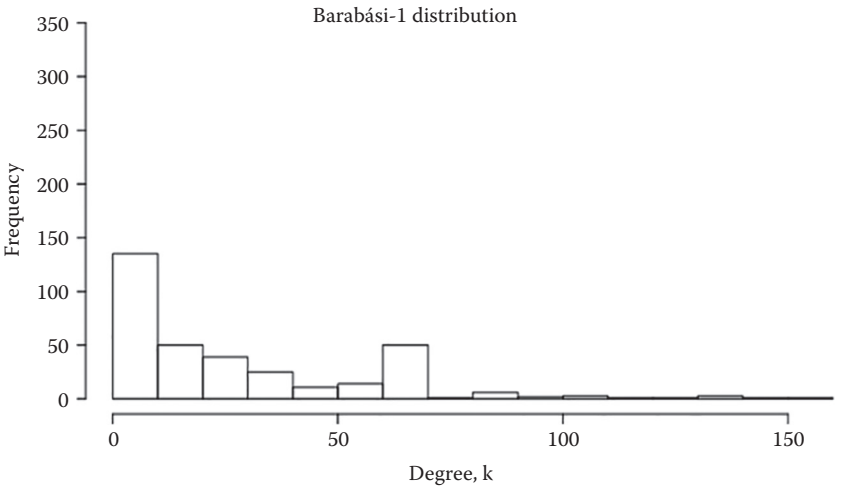


Figure 2.7 Distribution of nodes in the Barabási collaboration network.

Table 2.1 Network Measures for Erdős and Barabási Collaboration Networks

Summary Statistic	*Erdős-1*	*Barabási-1*
Node count	474	343
Edge count	1604	4883
Density	0.0143	0.08315
Mean degree	6.76	28.4
Max degree	51	154
Mean distance	3.8435	2.9268
Diameter	10	8
Centralization	0.093	0.3671
Largest component	98%	96%
Clustering coefficient	0.219	0.714
Assortativity	0.177	0.249

or dark portions of networks) that is still not well understood. Complex phenomena remain ill-understood because the traditional methods of data and mathematical analysis are insufficient to overcome that barrier of complexity. One way to advance that understanding is with better clustering of datasets.

Clustering Methods and Metrics

Clustering is a DA technique that can be applied to any collection of data with the goal of organizing them into clusters—subsets that have similar characteristics. The underlying principle is to define a distance measure between pairs of records and then partition the records for which the distances between pairs of records in the same cluster are small, and the distances between pairs of records in different clusters is large (Bertsimas et al., 2016).

In many applications, distance is determined by differences of values in a set of selected attributes. This permits weighted attributes reflecting the relative importance of the difference in each selected attribute. Part of the art of using this method is to define distances in a way that the resulting clusters have the desired property—data in each cluster have similar values for the selected attributes and data in different clusters have different values. Often clustering is the first step in data analysis. There are no universally accepted metrics used to evaluate clustering methods. There may be different objectives for different uses of clustering. In some cases, the user wants to predefine the number of clusters to be formed; in other cases, this is left unspecified. In some cases, it is desired that every record belongs to one set in the cluster. In other applications, not all records need to belong to a cluster, and some records may be permitted to belong to more than one cluster.

We measure the effectiveness of a clustering method by computing a penalty p for each pair of records that are close to each other and in different clusters, and a penalty q for each pair of records different from each other but in the same cluster. The size of the penalty for a pair of records depends on how similar or different the records are according to the defined distance function. An optimal clustering method corresponding to a specified metric is one for which the sum of the penalties, over all pairs of records, is minimized. This methodology and algorithm are similar to the linear discriminant used to find attributes that separate the dataset into different kinds of objects where the attribute being considered is geographic location.

Conclusions

Network science and data analytics are playing important roles in developing cyber science and complexity science. Our work hopes to build a foundation for further development and the application of these concepts and methods to data-related problems in military-relevant areas, such as cyber and complex science.

References

Arney, C. (2016). Cyber modeling. *UMAP Journal*, 37, 93–97.

Arney, C. and Coronges, K. (2015). Categorical framework for complex organizational networks: Understanding the effects of types, size, layers, dynamics and dimensions. In G. Mangioni, F. Simini, S. Uzzo, and D. Wang (Eds.), *Complex Networks VI*, pp. 191–200. doi:10.1007/978-3-319-16112-9_19.

Bertsimas, D., O'Hair, A., and Pulleyblank, W. (2016). *The Analytics Edge*. Belmont, MA: Dynamic Ideas.

Brandes, U., Robins, G., McCranie, A., and Wasserman, S. (2013). What is network science? *Network Science*, 1, 1–15.

Brantly, A. (2017). Innovation and adaptation in jihadist digital security. *Survival: Global Politics and Strategy*, 59, 79–102.

Burnett, D., Beckman, R., and Davenport, T. (2014). *Submarine Cables: The Handbook of Law and Policy*. Boston, MA: Martinus Nijhoff Publishers.

Carter, K., Riordan, J., and Okhravi, H. (2014). A game theoretic approach to strategy determination for dynamic platform defenses. In *Proceedings of the First ACM Workshop on Moving Target Defense*, New York: ACM, pp. 21–30.

Delle Fave, F.M., Shieh, E., Jain, M., Jiang, A., Rosoff, H., Tambe, M., and Sullivan, J. (2015). Efficient solutions for joint activity based security games: Fast algorithms, results and a field experiment on a transit system. *Journal of Autonomous Agents and Multiagent Systems*, 29, 787–820.

Department of Defense (2013). *Joint Publication 3-12 (R) Cyberspace Operations*. Washington, DC: US Department of Defense.

Department of Defense (2015). *The DOD Cyber Strategy*. Washington, DC: US Department of Defense.

deSolla Price, D.J. (1986). *Little Science, Big Science… and Beyond.* New York: Columbia University Press.

Holtz, R. (2015). How many scientists does it take to write a paper? Apparently, thousands. *The Wall Street Journal*, August 9.

Ingols, K., Chu, M., Lippmann, R., Webster, S., and Boyer, S. (2009). Modeling modern network attacks and countermeasures using attack graphs. In *Proceedings of the 2009 Annual Computer Security Applications Conference*, Honolulu, HI, pp. 117–126.

Korolov, R., Lu, D., Wang, J., Zhou, G., Bonial, C., Voss, C., Kaplan, L., Wallace, W., Han, J., and Ji, H. (2016). On predicting social unrest using social media. In *ASONAM 2016*, San Francisco, CA, August 18–21, pp. 89–95.

Ledford, H. (2015). How to solve the world's biggest problems. *Nature*, September 21.

Mayhew, M., Atighetchi, M., Adler, A., and Greenstadt, R. (2015). Use of machine learning in big data analytics for insider threat detection. In *Proceedings of the MILCOM 2015 – 2015 IEEE Military Communications Conference*, Tampa, FL, pp. 915–922.

National Academies of Sciences, Engineering, and Medicine (2017). *Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions.* Washington, DC: National Academies Press. doi:10.17226/23670.

National Research Council (2005). *Network Science.* Washington, DC: National Academies Press.

Paxton, N., Moskowitz, I.S., Russell, S., and Hyden, P. (2014). Developing a network science based approach to cyber incident analysis. *Paper presented at NATO IST-122 Symposium on Cyber Security*, Talinn, Estonia.

Roehner, B.M. (2007). *Driving Forces in Physical, Biological, and Socio-economic Phenomena: Network Science Investigation of Social Bonds and Interactions.* Cambridge, UK: Cambridge University Press.

Scientific American Editors (2015). State of the world's science 2015: Big science, big challenges. *Scientific American*, 313, 34–35. doi:10.1038/scientificamerican1015-34.

Shah, A. (2014). Why is biodiversity important? Who cares? *Global Issues*, January 19. Accessed August 9, 2017. http://www.globalissues.org/article/170/why-is-biodiversity-important-who-cares.

Shieh, E., Jiang, A., Yadav, A., Varakantham, P., and Tambe, M. (2016). Extended study on addressing defender teamwork while accounting for uncertainty in attacker defender games using iterative dec-MDPs. *Multi-Agent and Grid Systems*, 11, 189–226.

Sinha, A., Nguyen, T., Kar, D., Brown, M., Tambe, M., and Jiang, A. (2015). From physical security to cyber security. *Journal of Cybersecurity*, 1, 19–35. doi:10.1093/cybsec/tyv007.

Weaver, W. (1948). Science and complexity. *American Scientist*, 36, 536–544.

West, B. (2015). *Fractional Calculus View of Complexity: Tomorrow's Science*. Boca Raton, FL: CRC Press.

8

Fan Engagement, Social Media, and Digital Marketing Analytics at Duke University

RYAN CRAIG

Contents

Relationships and awareness—successful marketing relies on the optimization of both. Even the most remarkable product, discount, or opportunity is destined to fail if nobody is acquainted with its existence. Likewise, even the most intriguing marketing campaigns are fated to go ignored if the messenger does not develop equity with the consumer. It is important to realize, though, that "success" is not solely reliant on the variables that exist on the consumer's side of the marketing equation. Marketers must gather the appropriate information and ascertain the ideal target market before even beginning the marketing campaign. A misaligned campaign wastes both financial and physical resources and will set growth back indefinitely.

This chapter discusses marketing practices, concepts, and applications rooted in "analytics"—a catchall term that increasingly borders on cliché—but a space that offers tremendous benefit when applied in strategic fashion. Analytics have historically been associated with inward-facing platforms, such as those that allow an organization to understand more about itself or its patrons; for coaches to understand more about their players, teams, or prospects; or for athletes to understand more about themselves. Duke University is utilizing analytics for many of those same reasons, from player development to segmented marketing, and in that sense, is less of an innovator than

an adopter—albeit a fairly early one in the college athletics space. One area in which Duke is on the leading edge of analytics within college athletics relates to its fan-facing, crowdsourced data visualization platform known as #DukeMBBStats.

Case Study: #DukeMBBStats Data Visualization Platform

Duke University began working on this project well before it was even aware it was doing so. In the early 2000s, Curtis Snyder, an avid Duke basketball enthusiast and veteran of college sports administration, took on a pet project to be completed on his own time: digitize and organize every Duke basketball stat on record into a multilayered, searchable database. From computer files to drawers of folders to microfilm, Snyder tracked down box scores from as early as the 1905–1906 basketball season and added those statistics to the collection. Several modifications and roughly 10 years later, Duke University has arrived at a truly unique intersection of historical data and modern fandom.

Previously, the database had functioned as a resource for the media and school officials as well as a destination for the dedicated core fans who sought story lines in stats long before this became a mainstream mindset. As Duke University Athletics leaders planned out the latest version, the goal for the application changed. Athletics leaders saw an opportunity for the platform to appeal to a wider audience, more specifically, the ever-expanding collection of fans—diehard and casual alike—that see and feel the game through numbers, comparisons, streaks, and records.

In addition, the project would adopt a "For Duke, by Duke" mantra. Instead of a singular entity providing the vision and executing the build, Duke students, professors, and alumni would become the bedrock of a product destined to enhance the experience of the wider Duke community, including its loyal fan base both domestically and around the globe. In response to the growing amount of data available, player tracking analytics would be included, as would algorithms that could scrape the updated number sets and automatically generate insights and unearth trends within seconds that would take months to brainstorm and research otherwise. Campus outlets would be given access to the data for use in projects that consistently require unique

and interesting data sets, and from those work streams, additional interactive visualizations will be built and added to the platform. In totality, the goal to create was a dynamic, developing statistical and visual repository.

A traditional, static box score helps you answer a question. For instance, how many points did Brandon Ingram have in the team's last game against North Carolina? It is a simple enough query. But how does that point total compare to what other freshman did against Duke's archrival? How many times had that mark been achieved before? Based on historical comparisons, does that performance portend good things for his and the team's future?

#DukeMBBStats begs you to ask more questions. Think YouTube, minus the ads and multibillion-dollar valuation. While #DukeMBBStats does not (yet) include an algorithm that suggests other stats you may enjoy as YouTube does with recommended videos, it does induce the same inquisitive mindset. While you may visit the page with only one idea, request, or bar room debate in mind, the sheer volume of information, coupled with logical organization and easy navigation, will invite you to explore many more. The platform will allow people to look up program records, individual statistics, team accomplishments, and opponent data. Does it feel like the team is rebounding better than in years past? Does the rotation seem shorter and do the starters' minutes seem untenable? In a few clicks, you will be able to see how the team's profile stacks up against years that have resulted in a conference or national championship. You will also be able to share the results you find with built-in social media functionality that makes use of the hashtag that was intentionally used in the name of the platform itself. After all, what good is your newfound knowledge without a community of people to share it with?

#DukeMBBStats serves two purposes—one internal to Duke and one felt beyond the walls of the campus in Durham, NC. Internally, a project like this helps bridge the gap that often exists between the academic entities within higher education and the institution's athletics arm. Outside of Duke, the core function is to bring the fans unique, compelling, interactive, and ever-evolving content and supply fans with a reason to consistently engage with GoDuke.com, the digital hub of Duke Athletics and an ecosystem that allows

the department to involve fans in a variety of cross-marketing and revenue-generating initiatives.

Of all of the adjectives used to describe #DukeMBBStats, let's take a minute to further examine one of them: "unique." In the case of this particular project, its "unique" label does not simply pertain to the content itself, but in the concept or genre. "Unique" is a hackneyed term whose meaning has been devalued by people that are actually discussing something that is one of a few, but at its core, something that is truly unique represents one of one.

What if you were the first to execute a raffle, the first to offer a "buy-one-get-one" deal? Someone had to be first and at that time, that offer was rightly labeled "unique." That is what Duke Athletics has accomplished with this project. The first flat-screen television made waves for the way it changed the manner in which society consumed its living room entertainment. But those waves were the kind a pebble makes in a pond compared to the oceanic rip current that was felt through civilization when the first television of any kind hit the market. Sure the flatter screen enhanced the experience, but the first television revolutionized entertainment as it was known at the time. Duke Athletics hopes that #DukeMBBStats provides a similar disruption with regard to the way college sports programs engage with fans and schools through stats.

Marketers need to timely adjust to the changing mentality and intelligence of fans. Supporters are smarter than ever and the relationship is no longer a one-way interaction whereby the school or franchise simply directs its patrons with respect to where to go and what to do. Consumers have more options and require a deeper connection in order to take action, spend money, or give of their time. It is incumbent on Duke Athletics as a department, therefore, to adjust to where and how people are living their lives. Content on any website needs to be mobile optimized for an increasingly on-the-go audience. Social media platforms emerge, fall off, and evolve, forcing institutions to cater messaging to the various spaces depending on the type of content and the target demographics.

In reality, fans are looking for one thing: effort. They want to know if a sport property has taken the time to provide them with something of value, understand them as people, and appreciate how and where they want to receive marketing messages. We all know what

a lack of effort feels like. Have you ever received an e-mail or letter with the wrong name in the salutation? What about the envelope meant for "the current owner" of your home, condo, or apartment? How likely are you to open, much less read and lend thought to, that correspondence? Likewise, if every communication came with a request, a deadline, or an overt purchase solicitation, it would be easy to see how those on the receiving end would tire of the endless money grabs.

You want your brand to be one predicated on relationships, trust, and personalization. Bring people into your program/team and treat them as a part of what makes it great, because that is what fans are. Without fans, there are no teams. Too often, enthusiasts are treated like charity dinner guests—sure, they are welcome to sit at the table, but only at a cost. Yes, you need to sell tickets, merchandise, and concessions to pay for the salaries and facilities that allow the department to run successfully, but that does not mean there should not be methods for people to fully and genuinely interact with the program in ways that are complimentary.

The fans of Duke are perhaps more ingrained in the identity of the school and its athletic program than most. After all, Duke is known for its "Cameron Crazies"—technically speaking, the passionate student section that populates the entire lower bowl of Cameron Indoor Stadium on the side opposite the team benches. But the personality of the "Crazies" has moved well beyond that 100 foot stretch of bleachers. It permeates the rest of the building on game days and the homes and watering holes of the millions that watch games from wherever they are in the United States or abroad. That type of passion, commitment, and support—the type that leads the student contingent to eschew the dorm room that is costing them tens of thousands of dollars per year for a fabric tent on a plot of land known only as Krzyzewskiville—needs to be matched by a commitment on the University's behalf. The fans have shown their interest, and now they should be rewarded.

If properly built, those relationships can help you reach marketing nirvana, the ideal scenario where your brand thrives in an environment exclusive of your team's record. As of the writing of this chapter, the Chicago Cubs have not won the World Series in more than a century, but the seats are often packed in Wrigley Field because the

games are a destination. The history and aura of the club and ballpark outweigh the subpar results that have been achieved between the ballpark field lines. Fans of the Cubs identify themselves not as winners or losers, but as Cubs fans. Sure, everyone loves to support a champion, but any team can sell tickets, jerseys, and hope during a winning streak or in the middle of the glory years. It is when losses start to outnumber wins that people truly decide whether they identify themselves as fans of your team or as passengers along for the ride while the getting is good. Relationships help convert the latter into the former.

Duke University is nearing a crossroads of sorts. The basketball program has been fortunate to call Mike Krzyzewski its head basketball coach for more than 35 years, with over 1000 wins and 5 national championships to date. But sooner than later, although some might hate to admit, Coach K. will no longer be roaming the sidelines in his suit and tie. Sure, he will be a part of the program forever, but when arguably the game's greatest coach steps aside, a regression to the mean is (if not a certainty) a possibility that needs to be acknowledged. Now more than ever, Duke Athletics needs people to understand and appreciate that Duke basketball, and inherently Duke Athletics, is about much more than wins or losses. It is a family that extends beyond the campus and into the minds, hearts, homes, computers, and smartphones of all of the fans that identify themselves with one of the most prestigious universities in the world. At that point, it is less about being a fan of winning and more about aligning yourself with the University and its student-athletes.

With that, let's revisit the ecosystem mentioned earlier in this chapter, which includes GoDuke.com and its social media presence, and the concept of "awareness" that was discussed at the beginning of the chapter. With a newly minted fan engagement platform that should hopefully help instill trust in fans, enhance the program brand, bolster relationships with those across campus and across the world, and make GoDuke.com a destination for the millions that call themselves Duke backers, Duke Athletics now has an audience that is both frequently present and willing to listen. From there, the opportunities to cross-market and promote the stories of other athletics teams and athletes and generate revenue through avenues like video subscriptions, e-commerce, sponsorship, and philanthropy are nearly limitless.

In order to capitalize on those opportunities, you need people in your environment long enough and often enough for them to see the banner advertisement that may rotate through the server or notice the e-mail sign-up link that will allow you to communicate more directly to them when they are not visiting the website. A discount on video subscriptions is only as good as the number of customers that see it. Think of it as the digital version of the "tree falling in the woods."

Your visitors are not the only ones to consider when bringing people to and keeping people in your ecosystem. As more advertisers move away from their fixation on "quantity metrics" like page views, key performance indicators (KPIs) like reach and time spent on page have become more central to the story sales teams are telling. A platform like #DukeMBBStats, and others like this platform, registers high marks in both KPI and page view realms. The amplifying powers of the baked-in social media components can disseminate the content among users and their network of friends, family, and colleagues far faster than anyone could utilizing traditional communication channels. Also, the "unique" content entices a lengthier stay on the webpage. The days of performance-related stats and analytics being used solely for player development and coaching are over. Fans want to be involved as well, and platforms like #DukeMBBStats bring that desire to life.

Fan Profiles

This chapter previously discussed the importance of effort and the current expectation from fans that fans be understood and appreciated. While #DukeMBBStats represents a real-life example of how that is accomplished through men's basketball, in order to scale that to the entire department, Duke is investing a great deal of time and energy into the next major emerging division in college athletics: data and analytics.

Professional sport teams, Fortune 500 companies, and behemoths like the government have been in the "big data" game for years, but for the first time, we are seeing college athletics put its hat in the ring. As pressure mounts to generate more and more revenue to help pay the increasing costs of scholarships, facilities, and recruiting, even the

most monetarily successful departments are forced to mine for additional sources of revenue.

For some athletics departments, that means expanding stadiums and adding premium seating. For others, it means raising ticket prices or establishing capital campaigns. And for others still, it means developing and enriching the fan base and, as a result, the number of prospective customers. While the first several options are certainly plausible and perhaps even necessary, this chapter will now take a deeper dive into that last area—cultivating the fan base.

There are two ways to increase the potential purchase power of fans: add more of them or learn more about the ones you already have. I would argue the best strategy involves a hybrid of the two. First you learn about whom your fans are, and then you go about supplementing your contact lists with people fitting that profile. Sure, from simply a "reach" perspective, anyone can be a fan. But through some digging, you can find people that are more likely to purchase tickets, come to a game, and buy merchandise. With staffing the ubiquitous issue that it seems to be in college athletics departments, making the most efficient use of your time is paramount. When you are in essence preselecting fans that are more likely to purchase, and concentrating your marketing efforts on them, you are allowing staff to work with a stacked deck. Fan profiles—a panoramic view of your customers and any of the touch points they have with your franchise or program—allow you to start shuffling through the cards.

The idea here is to have a gauge on the full breadth of someone's interaction with you and your team. You want to know what they have purchased, how much of it they have purchased, where and how they purchased it, and how many times they have done so. Knowing someone is a season ticket holder is important. But knowing that they have been a season ticket holder for 8 straight years, that they bought the tickets online the day they were available for the past five of those seasons, and that they have increased the number of tickets each of the last 2 years is even better.

Additional information, without exception, helps formulate your strategy. Have they changed locations within the stadium? How many sports are they buying tickets for? Have they purchased any merchandise or VIP packages in their time as a fan? How far are they driving, or flying, to attend the games? There are dozens of questions

you can ask about the same person and that can happen only when you examine the ticket purchase.

For a season ticket holder at Duke, analytics and sales leaders might also ask about whether they have purchased an auction item related to the team, donated money to become an "Iron Duke," downloaded the team's schedule to their computer or mobile device, purchased a subscription to the Blue Devil Network Plus platform, purchased a ticket miniplan, signed up for an e-mail newsletter, or partaken in any of a plethora of other existing associations with the athletics department. All of those questions are designed to do one thing: give athletics leaders the best idea possible of who fans are and how fans behave.

Remember the piece about the mislabeled e-mail and how unlikely you were to click through and read what it had to say? What about the opposite end of that spectrum? What if you were able to segment the population you were sending that e-mail to as a way of letting the consumer know "we are not going to bother you with something you are not interested in?" What if you could let the fan know that you have done your homework? Imagine how much more likely you would be to not only open the correspondence but also to read the marketing material. This is where fan profiles become the basis for your specialized, segmented marketing efforts.

With a thorough understanding of your fan base, you can begin to narrow the scope of your marketing campaigns. Gone are the days of the 90,000 person e-mail blast—the message is too generic and the audience too large. Almost by default you are going to include thousands of individuals that are at best disinterested or at worst annoyed by seeing that tactic employed at the expense of space in their inbox. By coagulating your fan profiles into a data warehouse that allows you to run reports and consolidate like-minded or behaviorally similar consumers into a variety of categories, you can more precisely cater the pitch to a fan base that increasingly craves and demands that specificity. Instead of alerting everyone that has ever given a dollar to Duke in any form or fashion that a football jersey is being auctioned off on GoDuke.com, why not limit that e-mail to people that have shown an affinity toward football?

Could you be leaving sales on the table? Yes, you likely are. But I would argue the cumulative effect of the "blanks" you would be firing would be far more detrimental to the department or franchise than

the loss of those dollars. You never want e-mails or phone calls coming from your organization to become white noise. If they do, you are going to cultivate a group of fans that tune out, instead of listen, by default. As a consequence, you are going to spend an inordinate amount of time on the phone with those customers, either handling complaints related to the parking pass they did not renew because it was in one of those e-mails they now always ignore or answering questions that would have been resolved on the landing page you disseminated that went overlooked since the reader was on autopilot and not actually taking in the information.

Try and limit contact to the most critical messages (e.g., season ticket renewal, changes in procedure, and pricing changes) or times when you feel like a conversion is most likely. The credibility and relationships you will build up will far outweigh the dollars "left on the table."

That credibility can manifest itself in a number of ways, including more thoughtful customer feedback. That, in turn, can help you establish value on subjects related to the game day experience or the program/team as a whole. Instead of merely receiving comments in the form of criticism, which is often when people feel moved to write or call in, customers can instead be asked to take part in a focus group, participate in a conjoint analysis, and fill out a survey or do a phone interview with a customer service representative. The best way to know what a fan prioritizes among ticket price, food, and parking is to ask, but fans will not answer if they do not have a relationship with the organization. This segmentation should also lead to more conversion success. Specification and conversion have a fairly linear relationship to each other. In taking a holistic look at the profile of a fan, you can begin to understand their habits and upsell them into areas they are likely to see a fit.

If someone, for instance, is an Iron Duke, a season ticket holder for lacrosse and has downloaded the lacrosse schedule into their mobile device, then this person has in essence raised their hand from the crowd and shouted, "I'm a lacrosse fan!" In that case, they are more likely to feel moved by a campaign advertising an autographed lacrosse helmet on the auction platform or a chance to watch replays of games they missed (or ones they wanted to see again) on their mobile device through Blue Devil Network Plus.

Someone that has not attended a game or shown interest in the program would see the same marketing e-mail as white noise. They might be more interested in soccer, field hockey, basketball, or one of Duke's other 26 sports, so a campaign for them should look much different.

In building the profiles, you are in essence allowing the fans to identify themselves. Instead of having to ask them which sports they like, or which areas of the department they desire to hear more about, you can make inferences by simply organizing the information they have volunteered throughout the years. In this way, the connection is both less intrusive and more constructive.

So far this chapter has discussed the potential uses for big data and analytics when it comes to customer information that is already available to you. But what about adding to your potential client base? A more complete understanding of your fans can also steer the process of data augmentation. Third-party vendors can assist with cleansing, deduplicating, and visualizing data. Oftentimes, these vendors also allow you to enrich the profiles by adding data to the blank spaces you come across during the build. This information can be acquired in several ways: it can be provided by your analytics vendor, purchased on your behalf by your analytics vendor, or purchased on your own through a separate external source.

Since you will likely be paying by volume, zeroing in on what and who you want is vital. A thorough grasp of the information you already have will help lend clarity to both of those segments. Ultimately, you want to try and build out groups for each of the areas you would like to target. For Duke, that could mean any one of a number of clusters: potential Iron Dukes, probable football season ticket holders, or prospective auction bidders. In time, you will come to understand, for instance, what the typical Duke football season ticket holder looks like. From there, you apply that same set of demographic information to the general population and retrieve thousands of accounts that fit that description. Narrowing your scope and marketing to a group that is more likely to engage shows the fans, even subconsciously, that you have put in the effort.

The additional accounts born of that augmentation process become your next set of leads for the development office, ticket sales team, and marketing department. It is the equivalent of replacing a blind date

with one that you have had a chance to see and learn more about. You still are not sure there will be a connection, but you have a whole lot more information going into the first meeting than you would have had otherwise. You can cater your message to them more personally and both the customer service representative and the consumer can feel like they are talking more with a casual acquaintance than a complete stranger.

Sometimes, it is easier and more fruitful to simply keep the people you have instead of trying to recruit others. People are inherently less likely to give up something they already own than purchase something they have lived without until that point. In behavioral science, this concept is known as "loss aversion." In that way, reducing churn, whether it is related to ticket sales, donations, subscriptions, or any other rollover revenue streams, can be as valuable an exercise, if not more so, than going through the process of generating new customers. Obviously, you would want the two streams to be working in parallel with each other, but it is important to make sure the focus does not shift too much into generating new revenue and not enough into keeping what is already there.

Just like you can narrow down your marketing campaign to fans that fit a certain profile and you can seek to augment your contacts with similar individuals, you can also use your analytics to hone in on the inverse. While certain characteristics and buying habits would lead you to believe someone is more likely to purchase, renew, or engage, other traits can paint the picture of someone that is likely to churn, cancel, or leave. You should be putting as much time into the latter as the former.

As frustrating as it may sound, in many cases, new money often only serves to balance out money lost through churn. You can drive $1 million of new revenue, but if you lost $1 million worth of season tickets from people that did not renew, what have you really gained? Sure, you have not lost ground, but you also have not grown your business at all. The most efficient organizations work on both ends of the spectrum simultaneously.

Once you have a better idea of who that group of people might be, you can "randomly select" them for seat upgrades, VIP events, or other exclusive experiential perks to help transform them from someone riding the fence about their renewal into a long-term stakeholder.

Summary

The applications of data and analytics to college sports are a lot like your experience with this book. You have chosen to read this book. If you were not interested in the general topic of data and analytics, you would not have picked it up. But how did you come across it?

If someone had shoved a book in front of you about a subject you had no interest in, you would likely never get past the cover and have an impolite thought about the person who forced you to look at it in the first place. But that is just one barrier to entry.

Even if you did choose it yourself, there would have to be something to keep you rifling through the pages. Just because you like a genre or topic does not mean you will like every illustration of it. Every action movie aficionado has rolled his or her eyes at an unrealistic car chase. All sports fans occasionally turn away from a game because the scoring margin or pace is unappealing.

It is all about finding an interested audience and keeping their attention—in the end that is how you build trust. If you like an actor and have seen them in several movies, you are more likely to give another movie they are starring in a shot, even if you are not sure what it is about. That is because they have built credibility with you and you are willing to invest in something based on that connection. If you have enjoyed this book and this specific chapter, the next time you see the author on a panel at a conference or witness their name in a byline, you will be more likely to check in on what he or she has to say.

None of that has a chance at happening, though, if someone or something had not tipped you off to this book's existence. There was a moment in time when you had to make the decision to obtain a copy, but that moment could not have happened if you did not even know there was something out there to obtain.

From books to tickets to social media, your chances of having a long-lasting and consistently positive interaction with your constituents dramatically increase if you put in the effort to cultivate a relationship, offer something unique, and make the world aware it is there.

Hopefully, this chapter has served to check all three of those boxes.

1

Big Data and Analytics in Government Organizations

A Knowledge-Based Perspective

MATTHEW CHEGUS

Contents

Introduction

Big Data Analytics (BDA) has been a popular topic in the private sector for some time. However, less is understood about its application in the public sector. With increasingly knowledge-based services dominating the economy, the cultivation and deployment of various forms of knowledge and the tools that enable it are critical for any organization seeking to perform well (Chong, Salleh, Noh Syed Ahmad, & Syed Omar Sharifuddin, 2011; Harvey, Skelcher, Spencer, Jas, & Walshe, 2010; Rashman, Withers, & Hartley, 2009; Richards & Duxbury, 2015). Private firms often acknowledge the impact of organizational knowledge on innovation and firm performance (Walker, Brewer, Boyne, & Avellaneda, 2011). Yet, findings

related to knowledge management (KM) in public-sector organizations have been somewhat mixed (Choi & Chandler, 2015; Kennedy & Burford, 2013; Massingham, 2014; Rashman et al., 2009). Some draw parallels between private and public organizations, where both maybe delivering some type of service (Choi & Chandler, 2015), whereas others caution that the application of private-sector organizational knowledge frameworks to public bodies might be untenable due to the differences in organizational environments such as ownership and control (Pokharel & Hult, 2010; Rashman et al., 2009; Riege & Lindsay, 2006; Willem & Buelens, 2007).

Furthermore, just as theoretical insights differ, the use of technologies and tools differs between public and private organizations. It has been argued that BDA initiatives in public-sector organizations are generally underutilized and the value returned is less than expected (Kim, Trimi, & Chung, 2014). Conflicting goals, changing leadership, stewardship of values, and challenges in measuring outcomes are all thought to constrain the use of BDA in public organizations (Joseph & Johnson, 2013; Kim et al., 2014; Washington, 2014).

Ultimately, public-sector organizations serve the people, and it is this ideological orientation and the ensuing stakeholder relationships that determine the appropriate use of BDA and delineate the differences in application from the private sector (Riege & Lindsay, 2006; Walker et al., 2011). The processes associated with BDA can be used to effectively manage knowledge and thus produce better program outcomes if employed not just to collect and store data but also to learn from these data to create meaning and insight. This article, therefore, is an exploration of the current literature on organizational knowledge and its related fields such as organizational learning (OL), in an effort to develop a conceptual framework for the successful application of BDA in the public sector. To do so, a literature review on KM, OL, and BDA was conducted to identify current thinking related to the public-sector context. This document briefly defines the literature search, explores concepts of knowledge as it relates to public-sector organizational conceptual framework, and then discusses the framework developed based on the findings from the literature review. A series of propositions based on the conceptual framework is then provided.

Literature Search

A systematic literature search was conducted in combination with more directed literature reviews. We started with seminal works in KM to provide initial direction and insight and then conducted multiple searches of the recent literature through the Web of Science citation database and ABI/INFORM Global with key words relating to KM, OL, information technology (IT), and BDA. Three questions drove the search for current literature pertinent to a discussion on KM and BDA in the public sector: What are the key elements of effective KM in the public sector? What differentiates use of BDA in public organizations from that in private firms? How can public-sector organizations effectively manage knowledge supported by BDA? The initial searches, along with their search terms and findings, are described in the appendix.

Managing Knowledge: Organizational Knowledge and Learning

To better define a conceptual framework for BDA, it makes sense to first address the concept of organizational knowledge and learning. The use of knowledge in the organization is generally related to helping individuals and organizations learn, and the hierarchy of data, information, and knowledge is a well-discussed notion. However, the literature review suggests that the strict separation between data, information, and knowledge might not, in fact, be entirely appropriate to the ways in which organizations use knowledge.

Authors such as Polanyi, Dewey, Penrose, and Hayek have contributed to different theoretical perspectives of knowledge (Rashman et al., 2009). Nonaka and Takeuchi (1995), Tsoukas and Vladimirou (2001), and others have extended such insights by exploring conceptual models of knowledge within organizations. A core theme, discussed extensively by Nonaka and Takeuchi (1995), is the distinction between tacit and explicit forms of knowledge. Within this conceptualization, data would be considered explicit: it describes the specific circumstances of the moment and so maybe more easily measured and recorded through concrete means. From a constructivist perspective, knowledge, being inherently more generalized, is more abstract and subject to all manner of individual perception. However, Nonaka

and Takeuchi argue that such distinctions between explicit and tacit knowledge maybe a false dichotomy; the more generalized form may not exist without the specifics from which those generalized patterns were abstracted.

Data may thus be seen as the lowest level of *informational units* comprising an ordered sequence of items that becomes *information* when the units are organized in some context-based format. That is, information emerges when data items are generalized from a specific context such as an organizational problem or opportunity. Knowledge has been represented as the ability to draw distinctions and judgments based on an appreciation of context, theory, or both (Tsoukas & Vladimirou, 2001). More particularly, organizational knowledge would be created through a process of cognitive assimilation where decision makers consider information abstracted from a specific context (Richards & Duxbury, 2015), leading to an understanding of the current situation and the organizational response required (Tsoukas & Vladimirou, 2001).

The putative relationship between data, information, and knowledge appears to be that knowledge is built upon contextualized information units lower in the hierarchy. That is, the knowledge creation process is sequential, starting with data as its lowest level. At each subsequent level, individuals attempt to generalize in order to gain context-specific insight. This process of generalization is helpful as it allows information to be utilized in many more circumstances, patterns to be seen between divergent applications, and lessons to be learned from a variety of experiences. However, generalization may also be problematic. Generalization from specifics may seem relatively straightforward, but such conclusions maybe difficult to apply to other specific circumstances if overgeneralized or oversimplified, or otherwise, inappropriate inferences are made. Tsoukas and Vladimirou (2001) caution that individuals understand generalizations only through connecting them to particular circumstances. Fowler and Pryke (2003) raise a similar alarm, noting that, as discussed previously, knowledge is not just objective information but also the perception arising through each persons' experiences. Thus, a tension maybe seen between the specific form of information (data) and the more generalized form of information (knowledge) that gives credence to the notion that there is some kind of information flow between apparently distinct categories of knowledge.

This paper not only recognizes that different forms of knowledge are related but also supports Nonaka and Takeuchi's idea that such distinctions maybe false dichotomies. Specifically, this paper asserts that the only meaningful distinction between data, information, and knowledge is the level of generalization. The current notions of explicit knowledge exist as observable artifacts (such as a direct empirical measurement), whereas tacit knowledge is generated through the abstract process of cognitive assimilation. This reasoning leads to the model shown in Figure 1.1, where dimensions of knowledge range from low level (data) to high level (knowledge). However, how one may abstract knowledge from data is the resulting question of this assertion.

Pokharel and Hult (2010) describe learning as acquiring and interpreting information to create meaning. Indeed, other authors share similar sentiments. Barette, Lemyre, Corneil, and Beauregard (2012) described different schools of thought from cognitive-based learning to social constructivist learning; the former is characterized as changes in information based on reflections of individuals, whereas the latter is more the result of multiple people sharing their specific experiences and extracting commonalities. All three perspectives relate specifics to generalities through some sort of process or transformation indicative of Richards and Duxbury's *assimilation*. Barette et al. (2012) reflect this notion by saying "Knowledge management and OL models overlap in terms of common fundamental concepts related to learning" (p. 138). Fowler and Pryke (2003), Chawla and

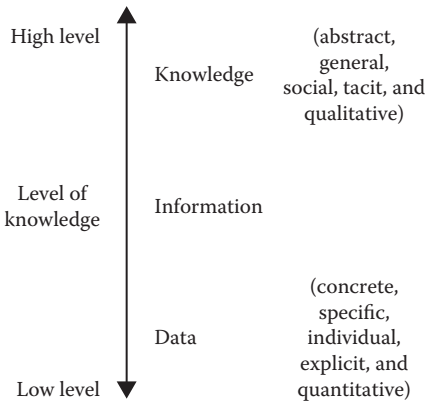


Figure 1.1 Dimensions of knowledge.

Joshi (2011), Kennedy and Burford (2013), and Harvey et al. (2010) echo similar observations. Learning, therefore, maybe considered the process by which information is generalized and abstracted to produce knowledge transitioning from lower levels of data to higher levels of knowledge.

As individuals undergoing this process would be relying on their previously acquired information, the process of learning would necessarily be influenced by all the previously acquired information, making learning a highly subjective affair. What one might recognize as a pattern might only be so because of previous patterns observed, for example. This would imply that learning is highly path-dependent, tacit, and idiosyncratic: "knowledge is not just objective information, but also is as much about the perception arising when information is refracted through the individual's personal lens" (Fowler & Pryke, 2003). These learning idiosyncrasies support the existing notions of the subjectivity of knowledge such as in the social constructivist view.

The Public-Sector Context

A number of common themes appear in the literature that describe the differences between private- and public-sector organizations: political influence being a significant contributor to organizational decision making (Barette et al., 2012; Pokharel & Hult, 2010; Rashman et al., 2009; Willem & Buelens, 2007), differences in power and control structures (Pokharel & Hult, 2010; Rashman et al., 2009; Willem & Buelens, 2007), accountability and transparency (Barette et al., 2012; Choi & Chandler, 2015; Greiling & Halachmi, 2013; Pokharel & Hult, 2010; Rashman et al., 2009), non-market not-for-profit orientation (Barette et al., 2012; Choi & Chandler, 2015; Rashman et al., 2009; Riege & Lindsay, 2006; Walker et al., 2011), public organizations motivated by stakeholder versus shareholder priorities in private organizations (Cong & Pandya, 2003; Rashman et al., 2009; Riege & Lindsay, 2006), constraints on organizational structure (Choi & Chandler, 2015; Pokharel & Hult, 2010), organizational fragmentation (Barette et al., 2012), and ambiguity of goals (Choi & Chandler, 2015; Willem & Buelens, 2007). These differences between private and public organizations lend credence to the notion that public organizations are, on a fundamental level, subject

to different influences than private organizations, and therefore, the process of learning and knowledge creation might also differ.

However, there are also some similarities that draw attention (Choi & Chandler, 2015). Both private- and public-sector organizations deliver services, for example, that would seem to be a point of commonality. Willem and Buelens (2007) argue that publicness is not, in fact, a dichotomy: government institutions (i.e., public administration, taxation, and national defense), public-sector institutions (i.e., schools and hospitals), and state enterprises, all may have varying degrees of *publicness*. Attributes such as ownership, funding, control, interests, access to facilities, and agency are qualities that may influence the degree to which an organization is public or private (p. 584), as Figure 1.2 depicts.

With this continuum of publicness in mind, New Public Management (NPM) attempts to take the notion of similarities between private- and public-organizational outcomes one step further by assuming that public organizations can and should benefit from private-sector methodologies that emphasize market orientation over traditional notions of public management (Cong & Pandya, 2003; Walker et al., 2011). Such an orientation suggests that managing performance in the public sector should follow from private organizations. Essentially, NPM provides a test for the underlying notions of similarities and differences between organizational sectors, and it was tested by Walker et al. (2011). The authors found for public organizations that market orientation has the opposite effect for private and public organizations (p. 715). Just because both sectors provide services to customers does not mean that they are motivated by, perform in similar ways to, or are evaluated against the same ideals.

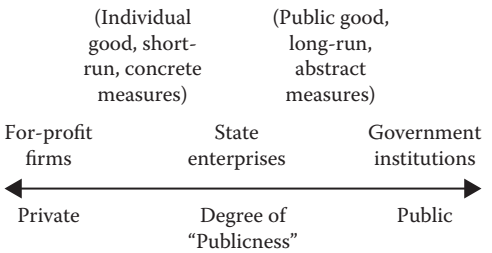


Figure 1.2 Degree of publicness.

Public- and private-sector organizations may face similar tasks. They may even produce similar outcomes and exist on a spectrum between private and public. However, fundamental differences in how these organizations perform, what drives their structures and decision making, and how they are judged to be successful suggest significant differences between private and public organizations. Such differences are large enough that the underlying assumptions of concepts like NPM should be put into question. A more general point that has started to emerge from this particular topic is the notion of a time horizon. Choi and Chandler's (2015) characterization of "myopic evaluation" (p. 144) implies an inappropriately short time horizon, which may not be comparable between sectors. Indeed, although one can argue that any organization that wishes continual existence should be concerned with long-run challenges, emphasis of private sector on quarterly results does not always reflect such a priority. With an assumption that a democratic system's public organizations exist to serve the public, especially in cases where the public good is best served by looking beyond the horizon of a single time period, much longer time horizons should be considered for all aspects of public organizations. The implication this has for KM is that public organizations tend to deal with higher levels of information and knowledge compared with private organizations because of their long-run outlook and broader scales of concern for the public good.

Consequently, not only the above-mentioned notions of dimensions of knowledge and publicness can be combined together, but also different organizational models maybe mapped to such a landscape. This landscape shows that private organizations tend to deal with lower levels of knowledge, are shorter in time horizon, and deal with more concrete measures of performance and accountability. By contrast, public organizations tend to deal with higher levels of knowledge, where more people are involved; time horizons are longer; and measures of performance and accountability are more abstract and difficult to define and measure. It must also be recognized that each organizational archetype would have many varieties, and so, there maybe examples of private organizations that deal with high-level knowledge and examples of public organizations that deal with low-level knowledge. For example, a corporate-planning exercise for a private multinational organization would necessarily include broader consideration

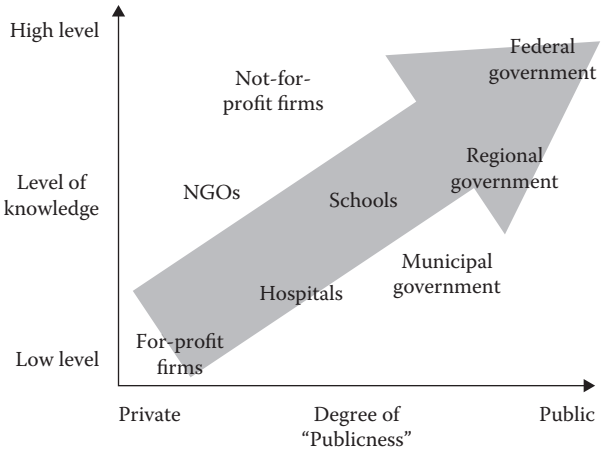


Figure 1.3 Conceptual framework-degree of publicness versus dimensions of knowledge.

than just a single individual's goals or just the next quarter's financial results, just as a municipal government maybe more concerned with local and immediate operational concerns, such as local infrastructure, whereas a federal government would be more concerned with long-term welfare of the entire population. Such propositions are reasonable from a level of analysis point of view; the grander the scale of people, time, and resources, the more general the inputs and outputs of those organizations, as measured by these qualities. Federal governments, for example, discuss ideological questions that explore how to best organize the country, whereas private firms discuss operational questions of how to exploit knowledge and resources for personal gain. Figure 1.3 provides the conceptual framework that maps the two dimensions of knowledge and publicness and places different types of organizations in relative position to each other.

The Role of Big Data and Analytics

Based on the conceptual framework, we can now explore the role of BDA within the knowledge-generating processes of public-sector organizations. Dixon, McGowan, and Cravens (2009) highlight the use of technology for KM in a public organization in two ways: to capture and to share knowledge (p. 256). Whereas data or information capture and dissemination maybe easily achieved, more abstract knowledge activities maybe more difficult. Advocates of technology

in KM describe a *coming of age* (Butler, Feller, Pope, Emerson, & Murphy, 2008) for the use of technology in knowledge creation and storage, retrieval, transfer, application, and administration (p. 262). O'Malley's (2014) account of a public-sector organization's adoption of Big Data seems to be quite positive based on its impact on performance: "we moved away from ideological, hierarchical, bureaucratic governing, and we moved toward information age governing-an administrative approach that is fundamentally entrepreneurial, collaborative, interactive, and performance driven" (p. 555). However, such a description seems to imply more data- and information-based processes that deal with the explicit component of knowledge.

Riemenschneider, Allen, Armstrong, and Reid (2010) argue that this situation might exist because decisions about technology in public-sector organizations are often crisis-driven and long-term planning is limited by political cycles. Accordingly, the focus of technology-based KM tools has often been on lower-level data capture and storage. Fowler and Pryke (2003) also note that the civil service is too narrowly focused on the management of explicit information. One might extrapolate this pattern of *data centrism* to conclude that most technological tools used in KM tend, particularly in the public sector, tend to deal well with data and information, as these informational units are more explicit and so more easily captured by IT systems. The abstract characteristics associated with knowledge mean that it is not as easily represented in these systems.

Kim et al. (2014) classify most current governmental applications of Big Data as at an early stage of development and are merely large traditional data sets that do not exploit the full potential of Big Data (p. 84). This is consistent with the development of analytic capabilities in organizations, which often begins with a data-centric approach such as investments in technology that help with the capture, storage, and transmission of information (Chen, Chiang, & Storey, 2012; Holsapple, Lee-Post, & Pakath, 2014). Moving beyond the data-processing stage, organizations start to derive benefits from data as they learn to better link their data sources to organizational context to create information and eventually knowledge. Context, as discussed previously, has unique characteristics in public-sector organizations. Going beyond simply capturing information, Joseph and Johnson (2013) describe the different types of analytics possible, such

as descriptive, predictive, and prescriptive analytics (p. 43), which can aid public organizations in the process of learning from data through reducing data complexity via generalization that provides a platform on which knowledge can be based.

Theorizing the Use of Big Data and Analytics in Public-Sector Organizations

The concepts discussed herein regarding KM, OL, and BDA in the public sector maybe summarized as follows. Knowledge maybe thought of as *levels of knowledge;* it begins at the lowest level of data and is abstracted or generalized through a process of learning to ever-higher levels of knowledge in a hierarchy where the former is the foundation for the latter. Low levels of knowledge are easily dealt with by IT tools, because they are more explicit and codifiable. High levels require more context and qualitative understanding in order to make sense of and use of such knowledge. These relationships constitute a dimension of knowledge.

Organizations may also be described along a dimension of *publicness*. Low-publicness (private) organizations have more individualistically defined scope based on a narrowed set of shareholders and often a shorter time horizon operating around more narrowly defined, and explicit, concepts of operational success. Highly public organizations, on the other hand, by definition, have broader scope based on the entirety of the public body that they represent and consequently have larger time horizons. Moreover, due to the conceptual and ideological nature of the highest levels of government, they have more abstract notions of success. Because of these differences, organizations that score higher along the publicness dimension will tend to operate on higher levels of knowledge as well. Higher levels of knowledge would thus require higher levels of learning from underlying data and sharing of that knowledge throughout the organization and so should act as a significant moderator for KM activities, including the technology used. Organizational learning's influence on the outcomes of organizational technology is partially supported by existing literature (Bhatt & Grover, 2005; Real, Leal, & Roldán, 2006; Tippins & Sohi, 2003), albeit from the private sector, suggesting that such a relationship maybe even stronger for public organizations.

From this high-level overview of the relationships between knowledge, learning, and organizational types, the following hypotheses are presented:

Proposition 1*: Higher publicness requires higher levels of organizational knowledge.*

Proposition 2*: The effectiveness of Big Data and Analytics in organizations that are highly public will be mediated by the level of organizational learning practiced in the organization.*

Proposition 3*: The degree to which organizational learning mediates the effectiveness of Big Data and Analytics that are highly public will scale with the degree to which that organization scores higher on the level of knowledge dimension.*

Proposition 4*: Big Data and Analytics will deliver more value in highly public organizations when combined with methods that enable reductions in data complexity, such as summarizations and visualizations, to enable rapid and effective high-level knowledge outputs.*

Proposition 5*: Big Data and Analytics will perform best in organizations that are highly public, when combined with other technologies and management practices that enable and encourage the rapid and continual sharing of organizational knowledge, particularly across organizational barriers.*

Discussion

Based on empirical research and the understanding afforded by convergent theoretical notions of OL and KM, the successful application of BDA in the public sector is expected to leverage OL to create and share high-level organizational knowledge within and beyond organizational barriers. However, this will not be an easy task. Many have suggested that public organizations do not easily facilitate the use of technology for high-level knowledge management due to their unique stakeholder environment. Top-down policy initiatives have largely failed to promote knowledge creation in public organizations, and organizational boundaries may fragment knowledge

(Rashman et al., 2009) and become an impediment to sharing (Fowler & Pryke, 2003; Massingham, 2014). In addition, political and bureaucratic power structures are not always aligned with the creation and proliferation of knowledge in public organizations (Girard & McIntyre, 2010; Joseph & Johnson, 2013; McCurdy, 2011; Piening, 2013; Willem & Buelens, 2007). Accordingly, although knowledge should be an important part of a public organization's operations, a number of significant barriers exist.

Jennings and Hall (2012) suggest a framework for identifying those organizations willing to support data and evidence-based decisions, whereby a low-conflict setting exists and the organization employs members with high scientific and technical capacity. However, the number of public organizations lacking political conflict, let alone engaging large proportions of scientifically and technically capable members, is likely to be low. Consequently, although BDA shows potential to enhance the high-level KM capabilities of public organizations, to do so would require the explicit direction and support from the many and varied stakeholders involved. Reaching consensus on these matters will likely happen first on the lower-level dimensions of knowledge, as such matters are more operational and short-termed, where the outcomes of increased knowledge capabilities can clearly be seen and argued through a business's value proposition. On the other hand, higher knowledge-based capabilities maybe contested for some time owing to political disagreements about organizational goals and ambiguity of the value of outcomes, which may not only be abstract in nature but also play out over longer time scales than a single election cycle. Consequently, Big Data and Analytics researchers and practitioners alike will have to take into consideration that the theoretical relationships between capabilities and outcomes will be potentially influenced by many intervening variables. Whether success is attainable will depend on the leadership, culture, and organizational structure necessary to support technological and learning activities. Time will tell how quickly BDA will proliferate into public organizations, but hopefully, these technologies will continue to provide enhanced abilities for the organizations that benefit everyone.

Appendix: Literature Review Search Terms and Findings

WEB OF SCIENCE	ABI/INFORM GLOBAL
KNOWLEDGE MANAGEMENT/ORGANIZATIONAL LEARNING	
((knowledge NEAR/1 manage) OR (organization* NEAR/1 learn*)) AND "public sector"*	*((knowledge NEAR/1 manage*) OR (organization* NEAR/1 learn*)) AND "public sector"*
Limit to 2010–2016 (inclusive)	Limited to 2010–2016 (inclusive)
Search in TOPIC (title, abstract, keywords, Keywords Plus)	Search in title OR abstract
Language = English	Language = English
Document type = Article, Review, Editorial	Limit to peer-reviewed and scholarly journals
BIG DATA IN THE PUBLIC SECTOR	
(big NEAR/1 data) AND ("public sector")	*(big NEAR/1 data) AND ("public sector")*
Limit to 2010–2016 (inclusive)	Limited to 2010–2016 (inclusive)
Search in TOPIC (title, abstract, keywords, Keywords Plus)	Search in title OR abstract
Language = English	Language = English
Document type = Article, Review, Editorial	Limit to peer reviewed and scholarly journals
ABSORPTIVE CAPACITY IN THE PUBLIC SECTOR	
((absorptive NEAR/1 capacit) AND "public sector")*	*((absorptive NEAR/1 capacit*) AND "public sector")*
Limit to 2010–2016 (inclusive)	Limited to 2010–2016 (inclusive)
Search in TOPIC (title, abstract, keywords, Keywords Plus)	Search in title OR abstract
Language = English	Language = English
Document type = Article, Review, Editorial	Limit to peer-reviewed and scholarly journals

Following the searches, a chronological review of articles from leading public-sector research journals was also conducted in the *Journal of Public Administration Research and Theory* and *Public Administration Review*. Articles were considered on their basis of overlap with concepts in knowledge management in the public sector. Once duplicated results were removed, a total of 178 articles were reviewed for their pertinence and excluded if not relevant. Finally, if there were topics that were deemed important for the issues at hand but were underrepresented in the resultant literature, additional articles were considered based on a specific search for those topics. The below-mentioned chart represents

both the included seminal works in knowledge management and related fields in addition to the most influential of the included search result articles that have formed the basis for the above discussion. The chart categorizes each article based on its application to the questions at hand and the contribution of each article to its respective area of study, in chronological order within each category.

ARTICLE	CONTRIBUTION
ORGANIZATIONAL KNOWLEDGE	
The knowledge creating company: How Japanese companies create the dynamics of innovation. (Nonaka & Takeuchi, 1995)	The creation of knowledge through a cycle (spiral) that is continuously changing form between tacit and explicit.
What is organizational knowledge? (Tsoukas & Vladimirou, 2001)	Individual knowledge becomes organizational knowledge through its codification and propositions underlain by collective understanding. Knowledge as the ability to draw distinctions and judgment based on context and/or theory.
Knowledge management in public service provision: The child support agency (Fowler & Pryke, 2003)	Empirically testing Nonaka and Takeuchi's model of five enabling factors for knowledge creation in a public-organization setting, finding tacit knowledge to be suboptimally managed in favor of information management.
ORGANIZATIONAL LEARNING	
How do public organizations learn? Bridging cultural and structural perspectives (Moynihan & Landuyt, 2009)	Learning as creating knowledge. Empirical test to find which variables foster organizational learning in a public organization: information systems, adequacy of resources, mission orientation, decision flexibility, and learning forums.
Organizational learning and knowledge in public service organizations: A systematic review of the literature (Rashman et al., 2009)	Data as ordered sequences of items; information as context-based arrangement of items. Organizational learning can be described as a process of individual and shared thought and action in an organizational context involving cognitive, social, behavioral, and technical elements. Social view treats learning as inseparable from social interaction. Knowledge is seen as a key component to learning, where knowledge is the content of learning.

(*Continued*)

ARTICLE	CONTRIBUTION
Varieties of organizational learning: Investigating learning in local level public sector organizations (Pokharel & Hult, 2010)	Learning involves acquiring, interpreting, and sharing information to create meaning and is a continuous process of knowledge integration. Individual learning feeds organizational learning. Public organizations may face more constraints to learning due to higher accountability expectations, increased stakeholder variety, and legal obligations in power and control structures.
Can government organizations learn and change? (McCurdy, 2011)	Public organizations that do not change tend to exploit pockets of political support that insulates them from change and perpetuates a lack of learning. Owing to this reluctance to change, most change may occur through *replacement*.
Dimensions of the learning organization in an Indian context (Awasthy & Gupta, 2012)	Test learning in a public organization with the Dimensions of the Learning Organization Questionnaire. Individual-level learning had positive effect on organizational outcomes when mediated by structural-level learning.
Organizational learning facilitators in the Canadian public sector (Barette et al., 2012)	Creation of a measurement instrument for learning in the public sector. Six main factors found are knowledge acquisition, learning support, learning culture, leadership of learning, strategic management, and the learning environment.
Accountability and organizational learning in the public sector (Greiling & Halachmi, 2013)	A narrow focus on short-term measures for accountability maybe inhibiting long-term organizational learning.
Exploration, exploitation, and public sector innovation: An organizational learning perspective for the public sector (Choi & Chandler, 2015)	Public organizations may lack appropriate feedbacks that would otherwise balance exploration and exploitation behaviors usually resulting from temporally myopic decisions.
Exploring the relationships between the learning organization and organizational performance (Pokharel & Choi, 2015)	Empirically testing seven dimensions of organizational learning in a public-sector organization. All seven dimensions showed positive relationship with performance. Organizational-level learning has a mediating effect on the relationships between individual and group-level learning and performance.

(*Continued*)

ARTICLE	CONTRIBUTION
CONTRASTING PUBLIC AND PRIVATE ORGANIZATIONS	
Issues of knowledge management in the public sector (Cong & Pandya, 2003)	Public organizations differ from private ones for two main reasons: public sector is stakeholder-dependent, whereas private sector is dependent on service delivery and is not threatened by survival.
Knowledge sharing in public sector organizations: The effect of organizational characteristics on interdepartmental knowledge sharing (Willem & Buelens, 2007)	An organization may be thought of in degrees of *publicness* rather than the traditional dichotomy based on ownership, funding, control, interests, access to facilities, and agency.
Impact of knowledge management on learning organization in Indian organizations—A comparison (Chawla & Joshi, 2011)	The impact of knowledge management on learning in vision, strategy, work practices, and information flow is found to be better for public organizations.
Market orientation and public service performance: New public management gone mad? (Walker et al., 2011)	Empirically testing New Public Management assumption that market orientation improves public service performance. Market orientation generally has a positive effect on consumer satisfaction but very little effect on organizational performance, which is the opposite of what is seen in private organizations.
A comparative analysis of conceptions of knowledge and learning in general and public sector literature 2000–2009 (Kennedy & Burford, 2013)	Schools of thought of knowledge management in the public sector lag behind more general knowledge literature traditionally aimed at more private organizations. Most existing literature on knowledge management in the public sector treats knowledge as static and codifiable, whereas contemporary scholars highlight the complexity of knowledge and its embeddedness.
KNOWLEDGE MANAGEMENT	
Designing a core IT artifact for knowledge management systems using participatory action research in a government and a non-government organization (Butler et al., 2008)	Advocates for tools/technologies as being vital for the execution of knowledge management. Specifically in the areas of knowledge creation and storage, retrieval/transfer/application, management, and system administration.
Knowledge sharing using codification and collaboration technologies to improve health care: Lessons from the public sector (Dixon et al., 2009)	Knowledge management is an evolutionary process that requires periodic evaluation and reflection in order to continuously improve quality.

(*Continued*)

ARTICLE	CONTRIBUTION
Knowledge management modeling in public sector organizations: A case study (Girard & McIntyre, 2010)	Introduces the Inukshuk model of knowledge management in the Canadian public service as each expression being unique and built upon a foundation of technology, culture, and leadership.
KM implementation in a public sector accounting organization: An empirical investigation (Chong et al., 2011)	Empirical test of a knowledge management framework in a public organization and its impact on performance. Knowledge sharing, technology, and leadership's impact on a knowledge-sharing culture are important factors.
An evaluation of knowledge management tools: Part 2—Managing knowledge flows and enablers (Massingham, 2014)	Empirical case study of knowledge management in a public organization. Highest-rated factor was knowledge preservation, and the most value was created through creating a *why context*, which gives meaning to information.
ENACTING KNOWLEDGE THROUGH CAPABILITIES	
Knowledge management in the public sector: Stakeholder partnerships in the public policy development (Riege & Lindsay, 2006)	Government functions, including policy, are based heavily on socially derived knowledge, which is difficult to capture. Effectively managed stakeholder relationships and the sharing of knowledge that results are integral to good policy. Such management needs to be considered dynamic.
Absorptive capacity in a non-market environment (Harvey et al., 2010)	A public organization maybe considered a knowledge-processing and -utilization entity, where the most important asset is the knowledge that is continuously renewed and created. Absorptive capacity occurs in three stages: exploratory learning, transformative learning, and exploitative learning. Absorptive capacity can both complement and integrate existing theories of knowledge management and knowledge processing in relation to performance.
Potential absorptive capacity of state IT departments: A comparison of perceptions of CIOS and IT managers (Riemenschneider et al., 2010)	Factors that affect a government IT department's absorptive capacity are creativity, innovative, and demonstrating initiative. It is also higher in departments that share information more readily. When an external environment is perceived as hostile, perspective of these departments will be one of reaction and minimization of risk taking.

(*Continued*)

ARTICLE	CONTRIBUTION
Evidence-based practice and the use of information in state agency decision making (Jennings & Hall, 2012)	Evidence-based decision-making capabilities in public organizations vary. Proposing a model to predict when a public organization will be evidence-based or not. Two dimensions: degree of conflict and degree of scientific capacity. Low-conflict, high scientific capacity will exhibit the highest levels of evidence-based decision making.
Written versus unwritten rules: The role of rule formalization in green tape (DeHart-Davis, Chen, & Little, 2013)	Formalized rules are often associated with organizational pathologies. However, rule-making capability, if use appropriately, can also increase effectiveness by improving rule design and compliance.
Dynamic capabilities in public organizations: A literature review and research agenda (Piening, 2013)	Dynamic capabilities maybe important for public organizations, which may face high rates of change due to frequent policy shifts. Development of dynamic capabilities follows three phases: learning through experimenting, enabling experimentation processes, and the management of ongoing tensions between innovation and exploitation. Management plays a key role in the facilitation of dynamic capabilities.
Knowledge sharing: What works and what doesn't work: A critical systems thinking perspective (Massingham, 2015)	The management of knowledge sharing should focus primarily on building social structures that can diffuse and embed tacit knowledge.
Work-group knowledge acquisition in knowledge intensive public-sector organizations: An exploratory study (Richards & Duxbury, 2015)	Information is data that has been organized to create meaning. Information that is assimilated is transformed into knowledge. Absorptive capacity is a form of knowledge acquisition in the Canadian public sector. Factors that positively affect absorptive capacity in public orgs are the role of managers, knowledge applicability, and the communality of knowledge for sharing.
BIG DATA	
5 keys to business analytics program success (Boyer, Harris, Green, Frank, & Van De Vanter, 2012)	Business analytics is a part of the whole organizational strategy, which should follow the business and not lead.
Big data and transformational government (Joseph & Johnson, 2013)	Barriers to government adoption of Big Data: Analysis of unstructured data, building Big Data infrastructure, acceptance of change in a highly bureaucratic environment, and data privacy.

(*Continued*)

ARTICLE	CONTRIBUTION
A unified foundation for business analytics (Holsapple et al., 2014)	Constructs an ontology of business analytics for further study. Provides a historical overview of analytical techniques in private-business organizations. Dimensions of analytics identified as domain, orientation, and technique. A general definition of analytics is proposed as "evidence-based problem recognition and solving that happen within the context of business situations."
Big-data applications in the government sector (Kim et al., 2014)	Big Data projects in public organizations are relatively immature. Success requires an ability to integrate and analyze information through new technologies, development of supporting systems, and the ability of Big Data to support decision making through analytics. Concerns of Big Data in government: security, speed, interoperability, analytics capabilities, and lack of competent professionals. A *business analytics framework* is proposed based on six building blocks: a *movement grounded in rationale*, a *capability set* of competencies, a *transforming process*, specific activities and practices, technologies, and the decisional paradigm under which evidence is evaluated and action is taken.
Big data and information processing in organizational decision processes (Kowalczyk & Buxmann, 2014)	Results from a multiple case study are presented. Data-centric approach is taken as Big Data addresses the supply of data. The *3-V* model of Big Data is introduced based on data volume, data velocity, and data variety. Organizational decision-making processes are discussed through information-processing theory, which has the goal of reducing uncertainty and equivocality through information processing as enabled by Big Data.
Doing what works: Governing in the age of big data (O'Malley, 2014)	Big Data is essential for transparency and accountability.
Big data and U.S. public policy (Stough & McBride, 2014)	Highlights that one of the biggest concerns of Big Data is the risk to privacy.
Government information policy in the era of big data (Washington, 2014)	Highlights limits of Big Data for use in transparency when in conflict with personal privacy.

References

Awasthy, R., & Gupta, R. K. (2012). Dimensions of the learning organization in an Indian context. *International Journal of Emerging Markets*, 7(3), 222–244.

Barette, J., Lemyre, L., Corneil, W., & Beauregard, N. (2012). Organizational learning facilitators in the Canadian public sector. *International Journal of Public Administration*, 35(2), 137–149.

Bhatt, G. D., & Grover, V. (2005). Types of information technology capabilities and their role in competitive advantage: An empirical study. *Journal of Management Information Systems*, 22(2), 253–277.

Boyer, J., Harris, T., Green, B., Frank, B., & Van De Vanter, K. (2012). *5 Keys to Business Analytics Program Success*. Big Sandy, TX: MC Press.

Butler, T., Feller, J., Pope, A., Emerson, B., & Murphy, C. (2008). Designing a core IT artefact for knowledge management systems using participatory action research in a government and a non-government organisation. *The Journal of Strategic Information Systems*, 17(4), 249–267.

Chawla, D., & Joshi, H. (2011). Impact of knowledge management on learning organization in Indian organizations—A comparison. *Knowledge and Process Management*, 18(4), 266–277.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.

Choi, T., & Chandler, S. M. (2015). Exploration, exploitation, and public sector innovation: An organizational learning perspective for the public sector. *Human Service Organizations: Management, Leadership & Governance*, 39(2), 139–151.

Chong, S. C., Salleh, K., Noh Syed Ahmad, S., & Syed Omar Sharifuddin, S. I. (2011). KM implementation in a public sector accounting organization: An empirical investigation. *Journal of Knowledge Management*, 15(3), 497–512.

Cong, X., & Pandya, K. V. (2003). Issues of knowledge management in the public sector. *Electronic Journal of Knowledge Management*, 1(2), 25–33.

DeHart-Davis, L., Chen, J., & Little, T. D. (2013). Written versus unwritten rules: The role of rule formalization in green tape. *International Public Management Journal*, 16(3), 331–356.

Dixon, B. E., McGowan, J. J., & Cravens, G. D. (2009). Knowledge sharing using codification and collaboration technologies to improve health care: Lessons from the public sector. *Knowledge Management Research & Practice*, 7(3), 249–259.

Fowler, A., & Pryke, J. (2003). Knowledge management in public service provision: The child support agency. *International Journal of Service Industry Management*, 14(3), 254–283.

Girard, J. P., & McIntyre, S. (2010). Knowledge management modeling in public sector organizations: A case study. *International Journal of Public Sector Management*, 23(1), 71–77.

Greiling, D., & Halachmi, A. (2013). Accountability and organizational learning in the public sector. *Public Performance & Management Review*, 36(3), 380–406.

Harvey, G., Skelcher, C., Spencer, E., Jas, P., & Walshe, K. (2010). Absorptive capacity in a non-market environment: A knowledge-based approach to analysing the performance of sector organizations. *Public Management Review*, 12(1), 77–97.

Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A unified foundation for business analytics. *Decision Support Systems*, 64, 130–141.

Jennings, E. T., & Hall, J. L. (2012). Evidence-based practice and the use of information in state agency decision making. *Journal of Public Administration Research and Theory*, 22(2), 245–266.

Joseph, R. C., & Johnson, N. A. (2013). Big data and transformational government. *IT Professional*, 15(6), 43–48.

Kennedy, M., & Burford, S. (2013). A comparative analysis of conceptions of knowledge and learning in general and public sector literature 2000–2009. *International Journal of Public Administration*, 36(3), 155–167.

Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-data applications in the government sector. *Communications of the ACM*, 57(3), 78–85.

Kowalczyk, D. W. I. M., & Buxmann, P. (2014). Big data and information processing in organizational decision processes. *Business & Information Systems Engineering*, 6(5), 267–278.

Massingham, P. (2014). An evaluation of knowledge management tools: Part 2–managing knowledge flows and enablers. *Journal of Knowledge Management*, 18(6), 1101–1126.

Massingham, P. (2015). Knowledge sharing: What works and what doesn't work: A critical systems thinking perspective. *Systemic Practice and Action Research*, 28(3), 197–228.

McCurdy, H. E. (2011). Can government organizations learn and change? *Public Administration Review*, 71(2), 316–319.

Moynihan, D. P., & Landuyt, N. (2009). How do public organizations learn? Bridging cultural and structural perspectives. *Public Administration Review*, 69(6), 1097–1105.

Nonaka, I., & Takeuchi, H. (1995). *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. New York, NY: Oxford University Press.

O'Malley, M. (2014). Doing what works: Governing in the age of big data. *Public Administration Review*, 74(5), 555–556.

Piening, E. P. (2013). Dynamic capabilities in public organizations: A literature review and research agenda. *Public Management Review*, 15(2), 209–245.

Pokharel, M. P., & Choi, S. O. (2015). Exploring the relationships between the learning organization and organizational performance. *Management Research Review*, 38(2), 126–148.

Pokharel, M. P., & Hult, K. M. (2010). Varieties of organizational learning: Investigating learning in local level public sector organizations. *Journal of Workplace Learning*, 22(4), 249–270.

Rashman, L., Withers, E., & Hartley, J. (2009). Organizational learning and knowledge in public service organizations: A systematic review of the literature. *International Journal of Management Reviews*, 11(4), 463–494.

Real, J. C., Leal, A., & Roldán, J. L. (2006). Information technology as a determinant of organizational learning and technological distinctive competencies. *Industrial Marketing Management*, 35(4), 505–521.

Richards, G. S., & Duxbury, L. (2015). Work-group knowledge acquisition in knowledge intensive public-sector organizations: An exploratory study. *Journal of Public Administration Research and Theory*, 25(4), 1247–1277.

Riege, A., & Lindsay, N. (2006). Knowledge management in the public sector: Stakeholder partnerships in the public policy development. *Journal of Knowledge Management*, 10(3), 24–39.

Riemenschneider, C. K., Allen, M. W., Armstrong, D. J., & Reid, M. F. (2010). Potential absorptive capacity of state IT departments: A comparison of perceptions of CIOs and IT managers. *Journal of Organizational Computing and Electronic Commerce*, 20(1), 68–90.

Stough, R., & McBride, D. (2014). Big data and US public policy. *Review of Policy Research*, 31(4), 339–342.

Tippins, M. J., & Sohi, R. S. (2003). IT competency and firm performance: Is organizational learning a missing link? *Strategic Management Journal*, 24(8), 745–761.

Tsoukas, H., & Vladimirou, E. (2001). What is organizational knowledge? *Journal of Management Studies*, 38(7), 973–993.

Walker, R. M., Brewer, G. A., Boyne, G. A., & Avellaneda, C. N. (2011). Market orientation and public service performance: New Public Management gone mad? *Public Administration Review*, 71(5), 707–717.

Washington, A. L. (2014). Government information policy in the era of big data. *Review of Policy Research*, 31(4), 319–325.

Willem, A., & Buelens, M. (2007). Knowledge sharing in public sector organizations: The effect of organizational characteristics on interdepartmental knowledge sharing. *Journal of Public Administration Research and Theory*, 17(4), 581–606.

Chapter 8

Big Data Analytics for Mobile App Security

Doina Caragea and Xinming Ou

Contents

This chapter describes mobile app security analysis, one of the new emerging cybersecurity issues with rapidly increasing requirements introduced by the predominant use of mobile devices in people's daily lives, and discusses how big data techniques such as machine learning (ML) can be leveraged for security analysis of mobile applications. In particular, the discussion focuses on malware detection for Android apps. ML is a promising approach in triaging app security analysis, in which it can leverage the big datasets in the app markets to learn a classifier, incorporating multiple features to separate apps that are more likely to be malicious from the benign ones. Recently, there have been several efforts focused on applying ML to app security analysis. However, there are still some significant challenges in making the solution practical, most of which are due to the unique operational constraints and the "big data" nature of the problem. This chapter systematically studies the impacts of these challenges as a set of questions and provides insights to the answers based on systematic experimentation results obtained from authors' past research. Meanwhile, this chapter also demonstrates the impact of some challenges on some existing machine learning-based approaches. The l arge (market-scale) dataset (benign and malicious apps) used in the above experiments represents the real-world Android app security analysis scale. This chapter is particularly written to encourage the practice of employing a better evaluation strategy and better designs of future machine learning-based approaches for Android malware detection.

8.1 Introduction to Mobile App Security Analysis

Mobile platforms such as Android are becoming the predominant computing utilities for end users. These platforms usually adopt an open-market model where developers submit applications to an "app store" for users to purchase and download to devices. The app store operators want to ensure that apps entering the market are trustworthy and free of malware. However, this is a non-trivial task due to the inherent undecidable nature of determining code behavior statically and the limitation of testing. Thus, app store operators adopt a variety of approaches to reduce the likelihood of "bad apps" entering the market and harming end users. This includes vetting of an app when it is first uploaded to the app store, and continuous vetting of apps that become popular. In addition, they constantly monitor issues reported by users, researchers, and companies to identify and remove malicious apps not flagged by the vetting process.

While app stores such as Google Play and Apple App Store have existed for years, current vetting technologies are still lagging behind the threats. This is evident from periodic reports of malware from these markets. The situation is worse in third-party markets. Even though the average malware rate for official markets like Google Play is low, with thousands of new apps uploaded to Google Play, new malicious apps are entering the official Google Play market without being detected

on a daily basis. While we did not find any official explanation from Google on why it has not done a better job at stopping malware, the scale of the app vetting process is clearly a factor. Early studies done by researchers showed that Google's app vetting service Bouncer only scans an app for 30 seconds each time to detect security problems [1].

While the extent of damage caused by those malicious apps is not clear, the possibility of them getting into app stores poses a non-trivial risk. Such risks need to be minimized by (1) curtailing the number of apps with security problems getting into the market and (2) quickly removing apps with security problems at the first sign. Both require effective analysis methods so that one can make quick and accurate decisions on which app has what security problems. This has to scale up to the large volumes of apps uploaded to app stores on a daily basis.

It is definitely not easy to achieve this. Common practice in industry, e.g., Google Bouncer and Amazon ATS, has adopted a variety of approaches including static and dynamic analysis. The research community has also designed advanced analysis methods and tools. But there needs to be an effective approach to address the *scale problem* in the vetting process. We observe that (1) although the number of apps in a market is huge, the number of malicious apps is not. If a "triage" process can effectively direct attention to the "right apps" for further examination, it could dramatically reduce the amount of compute and manual efforts; (2) the large number of apps in the markets actually provides an edge for defenders: it will allow us to identify patterns and trends that would be hard to find with smaller amounts of data [2]. The key to success is to identify, in an efficient and precise way, which apps are more likely to have security problems, so that the precious resources (human or computer) can be directed to those apps first. This triage problem has been examined in prior work [3] with promising results. The effect of big data in helping identify malware is further illustrated in the recent MassVet work [4], which aimed at quickly identifying malware created by repackaging existing legitimate apps with malicious payload. MassVet adopts a simple yet effective approach where an app is compared with a large number of existing apps in the market to identify "visually similar apps" with different components and "visually non-similar apps" with common components. The "DiffComm" analysis yields anomalous different or common components between apps which become the basics for identifying repackaged malware. This analysis can be done efficiently at a market scale. While these analysis techniques were invented to identify malware that are built in specific ways like repackaging existing popular apps, the threat landscape is certain to move toward more sophisticated malware creation processes that require more efforts from the malware writers, e.g., they may have to create their own apps that become popular instead of getting free rides on existing popular apps, or they may invent techniques to obfuscate the repackaging relations to break the assumptions of the specific detection techniques such as MassVet, and so on. This is inevitable given the rising stakes mobile devices bring to both individuals and organizations—mobile devices are now used for critical functions such as payments

and are becoming an integral part of organizations' enterprise IT systems. All this indicates that app vetting will be a highly complex and evolving process and it is not likely that a completely automated process without human intervention can do an adequate job. This is highlighted in Google's recent announcement to change its vetting process which now involves human review before an app can be published, instead of the completely automated process in the past. This puts more urgent need for better triaging capabilities since human labor is scarce given the amount of work needed, and expensive. An effective triaging can increase productivity by helping analysts to focus their effort on apps that are more likely to be malicious, and spending less time on those that are more likely to be benign. In the end, more general methods for triaging the vetting of apps on a large scale will be needed to address the evolving threats.

8.2 Applying Machine Learning (ML) in Triaging App Security Analysis

Machine learning is a promising approach in triaging app security analysis, in that it can leverage the big data in the app markets to learn a classifier, incorporating multiple app features to separate apps that are more likely to be malicious from the benign ones. Such separations are often times subtle and cannot be easily expressed by logical rules; machine learning is good at identifying hidden relationships in big data. A typical machine learning-based approach for Android malware app detection employs a classifier (e.g., an off-the-shelf machine learning classifier, such as k-NN) which is trained on a training set consisting of known benign apps and known malware apps. To evaluate the classification performance, the number of correctly and incorrectly classified apps is measured on a test set whose labels are unknown to the classifier at the time of evaluation.

Recently, there have been several efforts focused on applying machine learning to app security analysis [5–8]. However, there are still some significant challenges in effectively using a machine learning approach to triage mobile app security analysis, most of which are due to the unique operational constraints and the "big data" nature of the problem.

- *Noise and uncertainty on labels.* It is hard to obtain ground truths to train a machine-learned classifier for mobile app security. The degree of "truths" on the labels assigned to samples varies depending on the quality of information sources based on which the labels are assigned. The learning algorithm must account for this.
- *Imbalance on data.* The overwhelming majority of data samples for mobile apps are benign applications. The amount of malicious apps is minuscule

compared to the millions of good apps on the markets. This both presents a challenge in learning and puts a high requirement on the classifier's performance. For example, with a 0.1% malware prevalence, a 1% false positive rate would mean 10 times false alarms than true alarms on the market, clearly unacceptable in operations.

- *Feature limitation.* Features that can be extracted from an app in a computationally cheap way are weak indicators of security problems and many of them can be easily evaded by malware writers. To improve triage quality, a higher quality set of features is needed and more computation needs to be involved to derive features with more reliable attack semantics that cannot be easily evaded. This comes at odds with the scale challenge of the problem. In addition, the highly dynamic nature of adversarial behaviors means that predictive features *will change* over time. An effective triage must account for that and identify the optimal window for training.
- Although the results of the machine learning-inspired approaches look promising, many critical research questions still remain unanswered. There exists substantial room for clarification and improvement.

The above "big data" challenges result in additional concrete challenges when applying machine learning to Android malware detection, as described below:

Ensuring proper evaluation: These challenges arise in selecting the evaluation metrics as well as in collecting and preparing the data (e.g., correctly labeling the apps in training/test set). We see that in most of the current ML-approaches, (1) the evaluation strategy does not follow a common standard; and (2) the ground truth on which these approaches are evaluated lack reliability.

Algorithm design: These challenges arise in the design space of the machine learning approaches. One such challenge is to construct an informative feature set for the classifier. For example, in some works (e.g., [5]) the feature set contains hundreds of thousands of items, and many items (such as the names of the app components) are arbitrary strings at the app developer's choice. This raises a question on whether all items in this large feature set are really helping the classifier or if a subset can be sufficient (or even better).

The previously proposed ML-approaches focused more on a specific setting defined by factors such as specific evaluation metrics, ground truth quality, composition of the training/test data, the feature set, and others. The reported performance results are then measured in that particular setting. However, since the setting varies widely across different approaches, it is difficult (if not impossible) to fairly compare the results. For many of the recently proposed solutions, we are not aware of the impact of the above factors on the classifier performance.

8.3 The State-of-the-Art ML Approaches for Android Malware Detection

Drebin [5] works with a massive feature set (more than 500K features) containing different types of manifest features (permissions, etc.) and "code" features (URLs, APIs, etc.). Yet, Drebin authors demonstrated that the malware detection system is scalable, and it can even run on a phone in the order of seconds. Drebin's performance results are also very impressive.

DroidSIFT [8] is unique in designing features in terms of distance among API dependency graphs. It builds the API dependency graphs G for each app, and then constructs the feature vector of the app. The features represent the similarity of the graphs G with a reference database of graphs of known benign apps and malware apps. Finally, the feature vectors are used in anomaly or signature detection.

MAST [3] is a triage architecture whose goal is to spend more resources on apps that have a higher probability of being malicious, thereby reducing the average computation overhead for app vetting. This system utilizes a statistical method called multiple correspondence analysis (MCA). It uses permissions, intents, and the presence of native code to determine the probabilities of being malicious.

MUDFLOW [6] argues that the pattern of sensitive information flows in malware is statistically different from those in benign apps, which can be utilized for malware detection. From an app, it extracts the flow paths through static analysis, and these paths are then mapped to a feature vector that is used in a classifier.

8.4 Challenges in Applying ML for Android Malware Detection

Figure 8.1 illustrates the overall pipeline for using machine learning for mobile app security analysis. The large number of app samples goes through a labeling and feature extraction process. Then part of the data is used in constructing a machine-learned classifier, and the rest is used in evaluation. There are multiple challenges in each stage of the process.

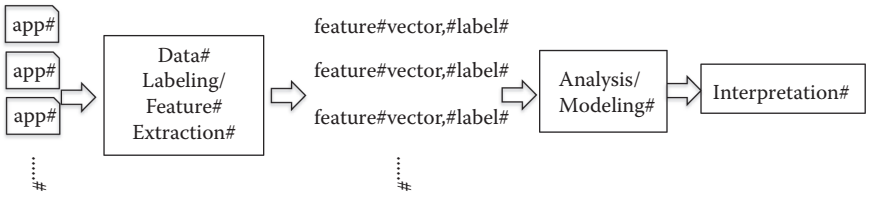


Figure 8.1 Big data analysis pipeline for mobile apps.

8.4.1 Challenges in Ensuring Proper Evaluation

To ensure that the ML-approach is evaluated properly is not straightforward. The related challenges fall under two subcategories as follows.

1. **Challenges in deciding the evaluation metrics.** The evaluation metrics for an ML-approach are not yet standardized and different ML-approaches rely on different metrics. For instance, DroidSIFT [8] and MUDFLOW [6] report the performance results in terms of true positive rate (TPR) and false-positive rate (FPR). Other existing works, such as MAST [3] and Drebin [5] present the receiver operating characteristic (ROC) plot, which is a generalized representation of TPR and FPR while the separating threshold is varied. Further, the ML-community has reported [9] that if the dataset is highly imbalanced, the PRC (precision-recall curve) is a better metric for measuring classifier performance than the traditional ROC curve. Given that the Android malware domain is highly imbalanced, i.e., the ratio of malware to benign apps in the real-world is highly skewed (1:100 or up), the above facts raise substantial doubt on whether current works are using the best metric.
2. **Challenges due to characteristics of the input data.** These challenges are related to data preparation, e.g., labeling the apps, composing the training/testing set, and so on. We see that these challenges are applicable to all the current ML-approaches. For instance, the age of input data may pose one challenge. Dated apps versus recent apps could lead to very different evaluation results in some cases. Deciding the data composition is another challenge, e.g., selecting the ratio between positive class (malware apps) size and negative class (benign apps) size in the test data, which may lead to different performance results of the classifier. To ensure realistic evaluation, we should conform to the real-life ratio of malware and good apps in the app store, but unfortunately this is not practiced in many existing works. Furthermore, the ground truth is noisy in reality while manually labeling a million plus apps is not feasible. So, we have to depend on security companies' reports on those apps (if available), which effectively lead to imperfect ground truth. We see that the ground truth on which the current ML-approaches depend is not fully reliable, which has a negative impact in two ways: (i) if training data has noise (mislabeled apps), the classifier mislearns things, which will negatively influence the classification performance. (ii) If test data has noise, we evaluate on the wrong ground truth, and then the reported performance results can be misleading. In addition, the presence of adware apps (which show unwanted advertisements to the user) in the dataset leads to further challenges. As adware has similarities to both benign and malware apps, it is often challenging to label an adware, e.g., including adware in the malware set or in the benign set, or dropping adware from the dataset altogether. The existing works differ on this choice, which further complicates attempting to compare their performance.

8.4.2 Challenges in the Algorithm Design

These challenges are related to the design of the ML approach itself. One challenge is to construct an informative (i.e., discriminative across the classes) feature set for the classifier. Some of the existing approaches are overwhelmed by this challenge. As an example, the Drebin approach [5] uses a very large feature set. One may want to know whether the classifier really needs this large feature set or only a subset of these items could be sufficient. We note that the size of Drebin's feature set is correlated with the size of its dataset—it has nearly 500K features while applied on the authors' dataset [5], but when we emulated Drebin feature's extraction on our larger dataset we achieved more than 1 million features. Do we really need these many features? How to identify and select strong, discriminative features is a challenge.

8.4.3 Challenges in Data Collection

We have discussed above the challenges due to characteristics of the dataset. Collecting a large dataset of apps poses a formidable challenge. Attempting to collect modern apps is an even more challenging task. Although Google Play provides the whole set of "free" apps (over 1.4 million), there is no "download API" available. So, we need to rely on app store crawlers like PlayDrone [10] that periodically scan the Google Play app store and collect entire snapshots of the store. The most recent apps, however, are not always available in the PlayDrone archive. Moreover, collecting a large set of adware and malware apps is also challenging—we have to rely on several sources. VirusShare and anti-virus companies provide large datasets of potentially malicious apps. These sets, however, are often noisy and impure, sometimes containing benign apps, Win32 binaries, and even blank apps. We believe that the large amount of data, even if somewhat noisy, provides further credibility to the results. To reduce the computation complexity of the machine learning approach is a further challenge. It is not straightforward how to design a scalable machine learning approach. When considering the millions of apps in the Play store, and the thousands of new apps added every day, scalability is of paramount importance. As an example of the degree of this challenge, we take note of MUDFLOW [6] authors' comment that sometimes their system took more than 24 hours to analyze one single Android app.

8.4.4 Insights Based on Our Own Study

Our research team has recently conducted an investigation of challenges that are faced in applying ML for Android security analysis [11]. We found that previously proposed machine learning approaches vary widely in terms of factors such as specific evaluation metrics, ground truth quality, composition of the training/test data,

the feature set, and others, making it difficult (if not impossible) to fairly compare the results. Some findings relevant to this chapter are listed below.

- *Is ROC the best metric to evaluate ML-based malware detection approaches?* The evaluation metrics for an ML-approach are not yet standardized and different ML-approaches rely on different metrics, such as true positive rate (TPR) and false-positive rate (FPR), receiver operating characteristic (ROC) plot, and the PRC (precision-recall curve). Given that the Android malware domain is highly imbalanced, i.e., the ratio of malware to benign apps in the real-world is highly skewed (1:100 or up), it is likely that PRC is a better metric for measuring classifier performance than the traditional ROC curve. Our investigation shows that indeed, the area under the PRC is a better metric for comparing results of different approaches in machine learning-based Android malware detection [11].
- *Does having dated malware in training/test set mislead performance?* The Genome Malware Project [12] has been used for many years as a main source of malware for many machine learning-based works. However, the Genome set, with malware from 2010–2012, has become a dated source of malware. We hypothesized that using dated malware sources together with more modern benign sources can lead to misleading results, and our study supports this hypothesis [11].
- *Does classifier performance degrade as we approach real-world ratio of malware and benign apps?* The occurrence of malware in the app stores is relatively low. This imbalance in the amount of malware and benign apps can contribute an interesting factor in classifier performance. Specifically, our results [11] show that the area under the PRC substantially degrades as the ratio increases (although the commonly used TPR and FPR do not change much), suggesting that results based on datasets that do not conform to the real data distribution could be misleading.
- *Does quality of ground truth affect the performance?* Peer works generally do some form of ground truth preparation for a machine learning classifier. Some works [5] require that a minimum of 20% of VirusTotal reports indicate the app is malicious. Other works [8] hold stringent standards and require that the reports return a matching malware family to be used in their dataset. In our own research [11], we investigated the effect of the quality of the ground truth data on the performance of the classifier, and found that the higher quality malware leads to substantially better results.

8.5 Recommendations

Below we present a number of recommendations for applying big data analysis to mobile apps. Some of them are specifically about the application of machine learning, while others involve complementary methods that could benefit the problem domain.

8.5.1 Data Preparation and Labeling

The community should explore and experiment with different ways to obtain ground truth information. In general, based on returned antivirus scanning results, we can separate Android samples into three categories:

1. *Ideal malware.* Apps in this category have highly credible labels and a high rate of shared labels among different antivirus companies, e.g., more than 25 different scanners are showing the sample as malicious and 20 of them give a shared keyword "DroidKungFu" in their scanning results. Thus, we can safely choose the shared label "DroidKungFu" as their family information.
2. *Candidate malware.* Apps in this category have either an unclear or a low rate of shared labels. For instance, only 10 out of 50 scanners identify the sample as malicious, even though 20 different scanners recognize it as malicious but only 5 scanners return a shared label. In either case, we cannot make a confident decision about the exact malware family, and only know that the sample is malicious.
3. *Unknown type.* Apps in this category do not have enough meaningful scanning results. The app could be benign but we are not really sure due to possible false negative in the antivirus products.

We expect the *ideal malware* datasets to be relatively small compared to the other two types of datasets, but cleaner. The *candidate malware* dataset is expected to be noisier, in the sense that we cannot label samples with high confidence. The *unknown type* dataset is the noisiest. By using such datasets, one can thoroughly study how a classifier's performance varies with the amount of noise.

Furthermore, given the uncertainty on the data label, it is interesting to study different approaches to labeling. For example, one can use the majority voting strategy to assign *hard* 0/1 labels to the samples. Alternatively, one can assign confidence scores to labels, based on the number of antivirus scanners that agree on that label, the trustworthiness of each scanner, and also the *freshness* of the application. By using information about "freshness," one can avoid a situation where all or most scanners identify an app as legit, and as a result the app will be labeled as benign with high confidence, when in fact it is a new type of malware. On the other hand, if a very small number of scanners identify an application as malware, while that application has been on the market for a long time, then there is a good chance that the application is legitimate. The confidence scores (which take values in between 0 and 1) associated with the two possible class labels of an app can be seen as *soft* labels (or probabilistic labels), and they essentially represent a probability distribution over labels for each instance. Intuitively, the soft labels can help capture (to some extent) the uncertainty on labels.

8.5.2 Learning from Large Data

To deal with large datasets, we recommend representing the classification problems at hand into a small hierarchy, where the problem at the root is the easiest, and has the largest amount of data, while the most specific problems—assigning malware to specific groups or categories—are the hardest, and have smaller amounts of data. More basic features can be used to address the more general problem (one advantage being that it will be less expensive to generate those features for a large number of apps), while semantics-richer features can be used for the more specific problems (such features will be more expensive to extract, but they will be generated for a smaller number of apps).

8.5.3 Imbalanced Data

Generally, the number of good Android apps is significantly larger than the number of malicious ones, leading to the challenge of learning from highly imbalanced data. We recommend the use of different standard strategies (such as under-sampling, over-sampling, cost-based learning, ensemble-based learning, etc.) to address the class imbalance problem.

8.5.4 Expensive Features

We envision systems that can help a human analyst in the vetting process. As part of this process, each new app will be classified using a hierarchy of classifiers. Furthermore, the classification process has to be fast. However, some classifiers (the most specific ones) may require expensive features whose construction can slow down the process, while a particular test app may be relatively easy to classify. To address this challenge and avoid generating expensive features unnecessarily, we suggest learning multiple classifiers for the same classification problem—from simpler classifiers that are based on more basic features to more complex classifiers that require more sophisticated features. Building a set of classifiers for each problem can be computationally expensive, but this task is done offline and, therefore, time is not the biggest concern (assuming the resources to perform the computation are available). At runtime, for a given app, we will first extract the most basic features, and use the classifier learned using those features to classify the app. If the app is classified with high confidence as benign or as malicious, nothing else needs to be done on the machine part. Otherwise, we will incrementally extract the next set of features required by the next classifier available for the problem at hand.

8.5.5 Leveraging Static Analysis in Feature Selection

An Android app is composed of one or more components and components interact (mostly) through mediated channels in the form of *intent*. While this nature

of Android apps makes static analysis challenging, it also creates an opportunity for presenting an app's behavior in a compact format that can help to extract rich features for machine learning.

As an example, consider a recent malware app called HijackRAT that we studied (illustrated in Figure 8.2). One of the malware's components—MyActivity—has the following behaviors: (1) By calling an Android system API, it attempts to hide the app's icon from the phone home screen and prevents the app from being stopped by garbage collection, and (2) it constructs an intent and sends it to DevicePolicyManager (a system service) to ask for administrative privilege. These two behaviors are suspicious and we can detect them by examining the code of this single component. We can detect behavior (1) by looking for the relevant API call in the code. We can detect behavior (2) by performing data flow analysis to resolve the target for the ICC call of the intent.

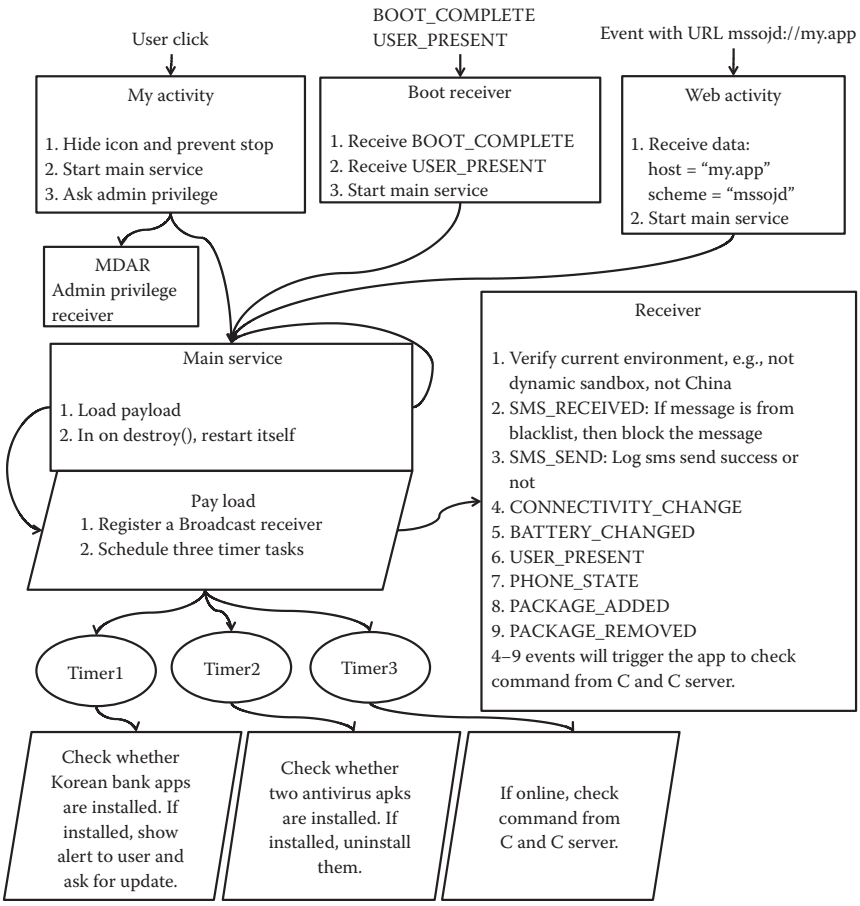


Figure 8.2 The ICC graph of malware HijackRAT.

On the other hand, this app also has inter-component behavior. After the component asks for admin privilege, it will save the user's decision (acceptance/rejection) to the SharedPreference (internal storage of an app). This will be retrieved by another component before it tries to perform actions that require admin privilege. If the user has not granted the privilege, the app will try to acquire it again. SharedPreference is a channel that the above two components use to communicate, which can be captured using static analysis.

MainService is the major component of the malware app. It dynamically loads a payload that is packaged in a separate file in the app's apk file. Upon running, the payload will register a broadcast receiver—a type of Android component that works like a mailbox receiving intents that can be filtered by it. Based on the type of the intent received, Receiver will perform various malicious functions, e.g., blocking messages from legitimate bank numbers so users are not aware of the nefarious transactions the app is trying to perform on the user's behalf. MainService will also initialize three times to perform other malicious functionalities.

A static analyzer like Amandroid [13] can detect the above behaviors and output them in the form of inter-component interaction graph (shorthand ICC graph) like the one shown in Figure 8.2. The unique advantage of this type of graph is that they provide a richer set of features that reveal an app's semantics in addition to other features such as API calls, source-to-sink flows, etc. Extracting a richer set of semantic features from an app is critical to the effectiveness of applying ML in triaging malware analysis, since malware writers can adapt and try to evade detection by changing the way the code is written. If features are based on code properties that can easily be changed, such as the choice of strings to name URLs or components, they will not be robust to evasion even if the classifier has very good performance results on the current malware data set. Features that are based on an app's behaviors are harder to evade since this would require the malware writer to change how the app achieves its objectives, which may only have a limited set of choices.

8.5.6 Understanding the Results

In addition to learning classifiers that can help in the malware triage process, one also needs to understand the results of the classifiers, especially to identify features that are predictive of problematic behaviors within an application. Information about predictive features can be used to inform how to better detect the problem using perhaps a slightly different static analysis plugin, and can help analysts in confirming/ruling out the results. A variety of methods can be used to perform feature ranking: wrapper methods, filter-based methods, and embedded methods [14]. Similar to learning classifiers from large Android app datasets, gaining insight into the results of the classifiers by performing feature ranking poses several challenges. Most importantly, the amount of labeled data available could be small for some classification tasks, while exhibiting high class imbalance. To address such challenges, methods for performing feature ranking from imbalanced data will be

beneficial, including semi-supervised/unsupervised methods. First, to address the imbalance challenge, one can use under-sampling, over-sampling, and ensemble-type methods to perform feature ranking [15]. As an example of the ensemble-type methods, one approach for learning from highly imbalanced data with an imbalanced ratio of 1:n works as follows. Construct n balanced subsets, where all subsets contain the same positive data (the minority class) and different subsets of non-overlapping negative data. Perform filter-based feature ranking on each subset and use the average scores to perform an overall ranking for the dataset. A similar approach, where the subsets could have overlapping negative data, was successfully used [16] on the problem of predicting software defects. Furthermore, semi-supervised-like approaches (e.g., transductive SVM) together with sampling approaches can be used to perform feature ranking using a recursive feature elimination-type algorithm [17].

8.6 Summary

In this chapter we discussed a number of challenges in applying big data analytic techniques, in particular machine learning, to mobile app security analysis. Many of the challenges are due to the scale of the mobile app market, e.g., Google Play. We present results from our own research that shows that consistent application of evaluation metrics in ML classifier performance is paramount to producing comparable results. The high imbalance in the positive and negative data samples in mobile data sets present unique challenges in both ML algorithm design and evaluation. We provide a few recommendations to approaches that can potentially address these challenges, and hope that they are useful for the research community to further research in this area.

References

1. Nicholas J. Percoco and Sean Schulte. Adventures in Bouncerland. *Black Hat USA*, 2012.
2. Alexandros Labrinidis and H. V. Jagadish. Challenges and opportunities with big data. *Proc. VLDB Endow.*, 5(12):2032–2033, August 2012.
3. Saurabh Chakradeo, Bradley Reaves, Patrick Traynor, and William Enck. MAST: Triage for market-scale mobile malware analysis. In *Proceedings of the Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec, pp. 13–24, 2013.
4. Chen Kai, Wang Peng, Lee Yeonjoon, Wang XiaoFeng, Zhang Nan, Huang Heqing, Zou Wei, and Liu Peng. Finding unknown malice in 10 seconds: Mass vetting for new threats at the Google-Play scale. In *Proceedings of the USENIX Security Symposium*, 2015.

5. Daniel Arp, Michael Spreitzenbarth, Malte Hubner, Hugo Gascon, and Konrad Rieck. Drebin: Effective and explainable detection of Android malware in your pocket. In *Proceedings of the NDSS*, 2014.
6. Vitalii Avdiienko, Konstantin Kuznetsov, Alessandra Gorla, Andreas Zeller, Steven Arzt, Siegfried Rasthofer, and Eric Bodden. Mining apps for abnormal usage of sensitive data. In *Proceedings of the ICSE*, 2015.
7. Hao Peng, Chris Gates, Bhaskar Sarma, Ninghui Li, Yuan Qi, Rahul Potharaju, Cristina Nita-Rotaru, and Ian Molloy. Using probabilistic generative models for ranking risks of Android apps. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security* (CCS'12), October 2012.
8. Mu Zhang, Yue Duan, Heng Yin, and Zhiruo Zhao. Semantics-aware Android malware classification using weighted contextual API dependency graphs. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security* (CCS), pp. 1105–1116, 2014.
9. J. Davis and M. Goadrich. The relationship between Precision-Recall and ROC curves. In *Proc. of the ICML*, 2006.
10. N. Viennot et al. A measurement study of Google Play. In *Proc. of the SIGMETRICS*, 2014.
11. S. Roy, J. DeLoach, Y. Li, N. Herndon, D. Caragea, X. Ou, V. P. Ranganath, H. Li, and N. Guevara. Experimental study with real-world data for android app security analysis using machine learning. In *Proceedings of the 2015 Annual Computer Security Applications Conference* (ACSAC 2015), Los Angeles, CA, 2015.
12. Yajin Zhou and Xuxian Jiang. Dissecting Android malware: Characterization and evolution. In *Proceedings of the IEEE SP*, 2012.
13. Fengguo Wei, Sankardas Roy, Xinming Ou, and Robby. Amandroid: A precise and general inter-component data flow analysis framework for security vetting of android apps. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security* (CCS'14), pp. 1329–1341, 2014.
14. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.
15. N. Chawla. Data mining for imbalanced datasets: An overview. In Oded Maimon and Lior Rokach, Eds., *Data Mining and Knowledge Discovery Handbook*, pp. 853–867. Springer US, 2005.
16. T.M. Khoshgoftaar, K. Gao, and J. Van Hulse. A novel feature selection technique for highly imbalanced data. In *Proceedings of the 2010 IEEE International Conference on Information Reuse and Integration* (IRI), pp. 80–85, Aug. 2010.
17. J. Weston and I. Guyon. Support vector machine—Recursive feature elimination (svmrfe), January 10, 2012. US Patent 8,095,483.

Chapter 1

Introduction: Big Data Analytics in a Learning Environment

Kristof De Witte and Jan Vanthienen

Contents

The abundance of data and the rise of new quantitative and statistical techniques have created a promising area: data analytics. This combination of a culture of data-driven decision making and techniques to include domain knowledge allows organizations to exploit big data analytics in their evaluation and decision processes. Also, in education and learning, big data analytics is being used to enhance the learning process, to evaluate efficiency, to improve feedback, and to enrich the learning experience. Before discussing some possibilities and issues in the use of learning analytics in education, we define its concept.

1.1 Data Analytics

With data available in large quantities, data analytics refers to a set of techniques and applications to explore, analyze, and visualize data from both internal and external sources. Applications can range from business intelligence (BI), enterprise reporting, and online analytical processing (OLAP) to more advanced forms of analytics, such as descriptive, predictive, and prescriptive analytics. Descriptive analytics identifies relationships in data, often with the intent to categorize data into groups. Predictive analytics exploits patterns found in historical data to make predictions about future behavior. Prescriptive analytics suggests decision options and shows the implications of alternative decisions.

Data can be of many types, and may originate from many sources: internal transaction data, web data, location data, browsing behavior, driving behavior, government data, and so on. With the increased digitization of society, a wealth of data is ready to be explored. The data explosion, however, has created a gap between the volume of data generated and stored, on the one hand, and the understanding and decision making based on these data, on the other hand. Traditional analysis techniques such as query and reporting or spreadsheet analysis are unable to cope with the complexity and size of current data sources.

This is where advanced analytics comes in, generating automatic *descriptions* of the data in terms of (human-interpretable) patterns, and *predictions* of unknown or future values of selected variables using, for example, clustering, classification, or regression techniques. In data analytics projects, however, a lot of effort is still devoted to collecting and integrating the available data; data preparation and cleaning; building, testing, and refining models; and finally, communicating results or triggering actions.

Data analytics has huge potential and is changing the world. The technical and managerial issues resulting from the adoption and application of data science in multiple areas are worth exploring (Baesens et al., 2016).

1.2 Data Analytics in Education

Rogge et al. (2017) argue that data analytics applications and functionalities provide a broad range of opportunities in the public sector. Their review reveals that governments worldwide have announced plans and road maps to support the development of big data in both the public and private sector. Education economists, in particular, are increasingly using the availability of large datasets (Rogge et al., 2017). As every step a student takes in the online world can be traced, analyzed, and used, there are plenty of opportunities to improve the learning process of students.

First, data analytics techniques can be used to enhance the student's learning process by providing real-time feedback, or by enriching the learning experience. The latter might take place in adaptive learning paths that provide a tailored learning environment for students. Thanks to the use of data analytics, the learning

environment can better correspond to students' characteristics in terms of cognitive abilities, earlier acquired knowledge and skills, interests, learning style, motivation or meta cognitive abilities. While similar adaptive and differentiated learning is difficult to realize in a physical classroom, it is relatively easy to realize in the online classroom. We discuss this more extensively in Chapters 2, 4, and 9.

Second, data analytics can be used to support the instructor or teacher. Using data analytics, the instructor can better trace, and take targeted actions to improve, the learning process of the student. By combining comprehensive student data with learning outcomes—in terms of student success, dropout, or cognitive skills—of earlier cohorts of students, the learning outcomes of the evaluated student can be predicted. Forewarned by similar indicators, instructors can pay additional attention to those students who are at risk of lagging behind. Moreover, the instructor can obtain descriptive analytics from the progress that students are making in online courses, or their use of tools in the electronic learning environment (see Chapter 2). In addition, the instructor can use data analytics to detect fraud by students (see Chapter 4) in a cost-effective way. Creating quality indicators (Chapters 7 and 9) for courses is also facilitated by data analytics.

Third, we see possibilities in using data analytics to measure the performance of instructors. Today, it is relatively difficult to compare and assess the performance of instructors. If the performance of instructors is measured by students' evaluations of teaching (SET), instructors with poor SET scores may argue that they face a more challenging student group, a more demanding topic, or that the students do not take the course seriously. Thanks to the abundance of data, these and similar arguments can be examined, and SET scores can be adjusted accordingly. De Witte and Rogge (2011) provide a model to do so. We will return to this issue in Chapter 5, where we discuss performance at the faculty level.

Finally, for policy makers, it is often unclear how schools use their available resources to "produce" outcomes. By combining structured and unstructured data from various sources, data analytics might provide a solution for governments that aim to monitor the performance of schools more closely. In Chapter 6, we discuss some techniques to relate resources to outputs (e.g., test scores, graduation scores). While similar techniques have existed for some time, school performance scores can now better capture the observed heterogeneity in school, student, and neighborhood characteristics, thanks to the increasing availability of data. This will facilitate the use of these performance scores for policy purposes.

1.3 How Has This Become Possible?

The upsurge in data analytics is a result of the automatic recording and ready availability of data in electronic form. The main enablers of this evolution are the availability of cheap data gathering, data storage, and computing technology. Information systems store and manage data about student background, registration,

program, and performance, and all these data can easily be exchanged, combined, and processed.

Moreover, the introduction of learning management systems and online learning applications allows huge amounts of data about the learning process to be collected in real time and at the source. Analogous to clickstream analysis or customer journey mapping in marketing domains, the availability of data about the learning process allows student behavior throughout the learning experience to be analyzed and described. This amount of education data, combined with student, course, and instructor information, makes it possible to use either traditional reporting techniques or more advanced forms of analytics, such as descriptive and predictive analytics.

1.4 Why Data Analytics Has Become Important

With the possibility of collecting and analyzing educational data comes the potential for enormous benefits through the proper use of data analytics. Policy makers (acting as principals in a classical principal–agent setting) are increasingly aware that, thanks to quantitative and qualitative data, they can better monitor the activities of the organizations they are funding (the agents). In fact, data analytics facilitates data-driven decision making such that, for instance, schools or universities can be better-compensated for their efforts in teaching students from disadvantaged backgrounds.

In addition, the availability of data allows stakeholders to assess the effectiveness (i.e., doing the right things) and efficiency (i.e., doing the thing right) of their interventions (see Chapter 10). While there is constant innovation in education (e.g., teachers who experiment with a different didactical instruction method), most of these interventions are not examined on their efficiency or effectiveness due to the absence of reliable data from before and after the implementation of the intervention. Data analytics can provide a solution.

At the same time, we need to be careful with the abundance of data. While combining data from various sources might create privacy issues, dealing with the overwhelming amount of data can also be an intricate issue (Chapter 9). It also forces us to think about such normative questions as whether we should store the data that are gathered during the student learning process. If we answer this question in the affirmative, to evaluate the effectiveness and efficiency of educational innovations, the question of how long we should store the data automatically arises.

Similar questions show that data analytics in education should not be the domain of a single discipline. Economists should discuss the possibilities, issues, and normative questions with a multidisciplinary team of pedagogists, philosophers, computer scientists, and sociologists. By bringing together various disciplines, a more comprehensive answer can be formulated to the challenges ahead.

This book provides a start to this discussion by highlighting some economic perspectives on the use of data analytics in education. We hope that the book marks the start of an interesting and multidisciplinary discussion such that, in the medium term, data analytics in education will seem as natural as a teacher in front of a classroom.

1.5 List of Contributions

This book on data analytics in education is structured in three distinct parts. The first section, consisting of three chapters, discusses the use of data analytics in to improve the student learning process. The second section, with four chapters, details the use of data analytics to measure the performance of faculty, schools, and students. In the third section, two chapters are devoted to the policy relevance of data analytics and the challenges ahead.

1.5.1 Data Analytics to Improve the Learning Process

Part I of the book begins with a chapter by Johannes De Smedt, Seppe vanden Broucke, Jan Vanthienen, and Kristof De Witte. The chapter focuses on supporting the automated feedback learning environment through process mining. It discusses some new ways to process student data, for example, by social network analysis. As similar data are shown to predict student performance, these can be used by instructors to obtain insights into student's behavior and to act accordingly in real time.

Chapter 3, by Wouter Schelfhout, provides a model, based on learning communities, as a platform for growing data use. Research indicates that data use by schools and teachers is not widespread, and where it does occur, is often superficial. In this chapter, we argue that schools and teachers are not open to data use because the essential conditions for integrating it in daily practice are not met in many schools. There is a profound lack of an effective professional development policy, which should start with the core processes and concerns of schools and teachers. Equally important is the frequently observed absence of shared instructional leadership as a basis for shaping this policy. Developing different forms of learning communities—in focused interactions—will provide a platform for addressing these challenges and needs, while at the same time promoting a gradual increase in the integrated use of data. Learning how to gather specific process data on teaching practices must form part of educators' professional development cycles to reach these goals. This will form a basis to give meaning to school internal output data and to school external data sources. Cooperation between schools and with external stakeholders such as education networks, governmental education departments,

and school inspectors will be needed to support this endeavor. As part of this contribution, a holistic model of "data for development" will be defined.

Chapter 4, by Silvester Draaijer and Chris van Klaveren, discusses the impact of fraudulent behavior and the use of learning analytics applications. Online quizzes are frequently used to prepare students for summative achievement tests. To encourage student participation, extra credits can be awarded to students who pass these quizzes. While anecdotal evidence indicates that offering quizzes carrying extra credit can result in fraudulent behavior in which students cheat to inflate their scores, there is as yet no empirical evidence investigating the extent of score inflation among, and its impact upon, cohorts of students. In this chapter, the impact of fraudulent behavior of first-year Dutch law students on weekly online quiz scores is studied. Exogenous variation in feedback to students was used to identify the impact. This exogenous variation was generated by abruptly, and without prior notice, ceasing to provide direct feedback to students on online quizzes. The main finding of the study was that the average quiz scores dropped by 1.5 points (on a scale of 0–10) immediately after the unanticipated feedback change. This result, first, supports the anecdotal evidence that online quizzes may not be a valid representation of student knowledge due to fraudulent student behavior. Second, and more importantly for this volume, fraudulent behavior may cause online quiz data to undermine the effectiveness of learning analytics applications.

1.5.2 Data Analytics to Measure Performance

Part II of the book focuses on the use of data analytics to measure performance.

Chapter 5, by Cristian Barra, Sergio Destefanis, Vania Sena, and Roberto Zotti, shows how data analytics can be used to disentangle faculty efficiency from student effort. In particular, this chapter provides an empirical methodology that allows monitoring of the performance of university students through data that are routinely produced and stored by universities. This approach disentangles the portion of the students' academic achievement controlled by the institutions' activities from the portion directly influenced by the students' own efforts, offering novel insights into the performance of universities and potentially supplementing the information from the standard league tables. The procedure is applied to a sample of 37,459 first-year students from a large university based in the South of Italy from the 2004–2005 to the 2010–2011 academic years. The evidence suggests that the efficiency with which faculties deliver their materials matters to female students and to students from low-income households. Pre-enrolment information such as the high school grade and type are good proxies of the student's own effort.

Chapter 6, by Maria C. Andrade e Silva and Ana Camanho, focuses on the measurement of performance at school level. In the majority of European countries, evaluation of schools is at the heart of the educational system as a means to guarantee the quality of education. Every year, in most countries around the world, students perform national exams. Their results are analyzed by several stakeholders,

including governmental agencies, the media, and researchers on educational issues. Current advances in information and communication technology (ICT) and data analysis techniques allow schools to make use of massive amounts of data in their daily management. This chapter focuses in particular on the use of students' data to benchmark schools. It illustrates the potential contribution of the information gathered and analyzed through data analytics to promote continuous improvement in schools' educational processes.

Chapter 7, by Kristof De Witte, Grazia Graziosi, and Joris Hindryckx, provides a methodology to handle the abundant data in the learning process. Using a unique and rich dataset of a large European higher vocational institute, this chapter examines which student, teacher, learning, and didactic characteristics can explain differences in students' performance. It proposes a two-step procedure, rooted in the data analytics literature, to select, in a data-driven way, the relevant variables from a large number of potential covariates. First, an importance measure of the variables is computed by repeatedly performing the Lasso variable selection technique on bootstrap subsamples of the original dataset. Second, the results of the bootstrap Lasso selection model are included as *a priori* information for a Bayesian factor approach. This allows the computation of the posterior probability distribution of different models, that is, the inclusion probabilities of potential variables. From this perspective, the chapter suggests an innovative approach to variable selection; by incorporating prior information into Bayesian analysis that is different from the standard choices (e.g., uniform distribution assumption).

Chapter 8, by Tommaso Agasisti, Francesca Ieva, Chiara Masci, Anna Maria Paganoni, and Mara Soncin, provides insights into using data from administrative sources. While public administrations collect data from various sources (e.g., test scores, income, taxes, juvenile offences, school history, medical records, and traffic violations), similar data are not combined in most countries (notable exceptions are the Netherlands and Hungary). Chapter 8 starts with a discussion on how administrative datasets are structured. It then provides some examples of well-known educational datasets: the PISA data, collected by the Organization for Economic Cooperation and Development, and the Italian Invalsi data. Chapter 8 concludes with some empirical applications of similar data to measure the performance of schools.

1.5.3 Policy Relevance and the Challenges Ahead

The third and final part of the book focuses on policy relevance. In Chapter 9, Kurt De Wit and Bruno Broucker define what is typically understood as "Big Data." They apply the concept to higher education, and focus on the data-related issues that need to be addressed. The authors specify what a governance structure for Big Data in higher education would entail. In particular, they focus on IT governance, internal governance, and external governance. Next, they illustrate what this can actually mean in practice by examining recent developments in Belgium (Flanders),

in Belgium (Flanders), at both the Flemish higher education system and higher education institution levels. This case is interesting because data-driven decision making, supported by interconnected databases and purpose-built analytical tools, has expanded at both levels in recent years.

The final chapter of the book, written by Wim Groot and Henriëtte Maassen van den Brink, makes the case for evidence-based education. Evidence-based education is the philosophy that education should be based on the best evidence about what works. This means that specific educational interventions, strategies, and policy science should be evaluated before being recommended or introduced on a wide scale. Without evidence, these interventions should be introduced on an experimental basis, such that their effects can be evaluated scientifically. The availability of large datasets may facilitate the paradigm of "evidence-based education."

References

Baesens B., Bapna R., Marsden J., Vanthienen J. and Zhao J. (2016). Transformational issues of big data and analytics in networked business. *MIS Quarterly*, 40(4), 1–12.

De Witte, K. and Rogge, N. (2011). Accounting for exogenous influences in performance evaluations of teachers. *Economics of Education Review*, 30(4), 641–653.

Rogge, N., Agasisti, T. and De Witte, K. (in press). Big data and the measurement of public organizations' performance and efficiency: The state-of-the-art. *Public Policy and Administration*.

Chapter 6

Big Data and the Coming Historical Revolution: From Black Boxes to Models

Ian Milligan and Robert Warren

Contents

Traditionally, historians have gone to archives—up in the morning, off to the airport, flying across the country, or even an ocean, and physically sitting in a reading room. These research trips imposed a substantial bottleneck on the amount of primary source research that historians could do. All of this has begun to change in the past three decades with the widespread advent of digitized primary sources, a force that historians are beginning to realize (or should be, in any event) is fundamentally transforming their research.[1–3] This is a force that touches all historians, be they those who use volumes of digitized primary sources or even just those who use databases or other repositories to navigate newspaper articles or other print

volumes. The new bottleneck increasingly relates to the consumption of this abundant information and performing analysis on it.* We used to be limited in the amount of time we could spend in an archive in Washington, DC, for example; now, we are limited in the amount of time we can spend sifting through all of this information online and making sense of it. New tools, methods, and scholarly frameworks are needed to deal with this material.

The shift toward online sources has been positive in many respects, particularly in the global reach it gives scholars, the lowered barriers to access, and the speed by which many historians can find the information they need. Yet it has also come with downsides: a lack of understanding of the Optical Character Recognition (OCR) algorithms that underpin source digitization, inattention to what has *not* been digitized, and, in some cases, the loss of context surrounding individual documents or collections. History, previously the province of painstaking archival work, is becoming a discipline dominated by website keyword searches. This chapter does not indict that process—indeed, we believe, in general, that the more accessible nature of historical research and new technologies has been a net positive—but rather argues that this shift requires a methodological rethinking. Engagement with this new digital world requires critical reflection and interdisciplinary engagement.

This chapter is organized as follows: we begin by reviewing the ongoing problems of data scale within the digital humanities. We then review different models of computational support for the digital scholar and different models of collaboration and publications for scholars. Given our experiences and situation as academics in North America, this largely draws on the professional experiences of that community of practice. Finally, we conclude with observations about the nature of scholarly work and its historic ability to adapt to new tools as they become available.

The Flood: From Close Reading to Black Boxes

As humanists, historians come from a tradition where each document has traditionally been understood through the lens of close reading.[4] This has led to a high degree of interaction between an individual scholar and his or her documents. Achieving this was costly: physical travel to archives and libraries to consult records. With the advent of online archives and databases, historians now have tremendous access to global resources, as well as the opportunity to discover their existence through search engines. Yet, this tremendous access to material has the consequence that the work of actually performing analysis takes up a larger proportion of time.

In short, physical access is no longer the defining factor for a scholar's research. The intimate research relationship with sources that scholars previously enjoyed is

* A shift effectively discussed in Roy Rosenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *The American Historical Review* 108, no. 3 (June 2003): 735–762, doi:10.1086/529596.

not necessarily possible at scale. While enormous amounts of gains have been done in the accessibility sense, much remains in the analysis and consumption sense.

This is intimately connected with the rise of the digital humanities as a field of scholarly analysis. Defining the digital humanities is not straightforward, as witnessed by the amount of discussion on that very topic.[5,6] Part of this field considers the data consumption, processing, and analytical and synthesis processes that are used by scholars in order to perform their research, which is intimately connected to the work we consider here. This thus lies at the heart of the paradox of the digital turn. It is now a victim of its own success: the information relevant to the scholar is now present in such quantity that it overwhelms the scholar's ability to consume it. This means that a critical scholarly infrastructure is needed to deal with scale, something that we need to tackle now.

The capacity to process historical information has not kept pace with the capacity to retain digital records. As an example, historians who have studied a small- to medium-sized event in the 20th century, such as the Canadian New Left, can honestly state that they have reviewed much of the extant formally archived primary documentation and drawn a conclusion from it.[7] Nothing is perfect, of course—sources have always been missed, interviews forgotten, memory changed, documents destroyed—but the scale at work meant that quite a bit of the preserved material was reviewed. Scholars reviewing the book would also be familiar with the largest collections and be able to explore their validity.

Scholars of digital era events, facing abundance, will be able to read a much smaller fraction of this information. A period where a social movement, such as Canada's First Nations Idle No More movement, can leave behind over 55,000 tweets in one single day suggests that the percentage of sources that a scholar can directly read has dwindled to a much smaller percentage. While scholarly access to such large repositories is not yet a settled question, the Library of Congress preserves these tweets; increasingly, individual researchers too are creating and storing their own curated data sets of tweets pertaining to large-scale events such as the Women's March, Presidential elections in the United States, and beyond.* We can see this in other events over the last 30 years: the First Gulf War, the e-mail records of the Clinton Administration, the World Trade Center bombings, or the events of the September 11, 2001, attacks.

Fears around digital abundance are not new. Chad Gaffield, a historian at the University of Ottawa, provided an overview of the field in a recent *Digital Studies* article. Indeed, he quotes the president of the American Historical Association Carl Bridenbaugh—*in 1963*—bemoaning the coming tide of data. "Among other ways," Bridenbaugh declared during his presidential address, "bigness has struck us by proliferating sources and editing, thereby deluging us with an overwhelming

* The Library of Congress announced that they were working with Twitter to preserve a digital archive of publicly accessible tweets in 2010. Access remains unclear. For a database of available Twitter datasets, see http://www.docnow.io/catalog/.

mass of data for the study of the last one and a half centuries of history."[8] Just as the problem is familiar, so too is the difficulty of exploring this abundance via computing. This quotation from a letter to the American Library Association (in 1962!) is oddly prophetic in the stresses between scholar, information tools and those that build the tools:

> Here is the basic weakness of information retrieval. It can only work with the values of the past. A computer cannot think. It can only remember what someone has told it to remember. Who is to decide what material is relevant to a subject? If he who programs the computer is as capable as the scholar in relating relevant material, he is wasting valuable time. He should be doing research.[9]

This highlights the inherent contradiction at play. On one hand, scholars want the scalability of computational processing while at the same time have concerns around losing control to a black box mechanism they do not understand. For example, when using Google, it may seem straightforward that the first page of responses for "Canadian history" include Wikipedia, the Government of Canada's citizenship guide, and some other projects; but why is the "First Peoples Historical Overview" or the history of the First World War Battle of the Somme relegated to the 10th page, where few will find it? This is an excellent example of how algorithms and computers begin to shape the work that we do.

What does it mean for a historian to turn analysis over to a computer? There is a precedent for this type of arrangement: previously, scholars would often make use of typists and typesetters to convert their manuscripts into publishable form. With the advent of affordable software applications, the difficulty of using computers for word processors has been greatly reduced and scholars now tend to do their own editing. Spreadsheets, for example, have greatly enhanced our ability to perform repetitive mathematical operations and basic statistical calculations.

Hence, the software application is now the mediating element between the computer and the scholar, but only within a limited context. Part of the promise of digital tools is to expand that context so that more questions can be answered by the scholars themselves, either through the use of prepackaged software or by having the scholars themselves write their own software.

Single-purpose, prepackaged software development is a complex endeavor that usually requires a team of specialists. This cost means that only the most generalizable problems are therefore tackled and that the more obscure computational needs of digital humanities need to be tackled by the scholars themselves. An additional complicating factor in the use of an external programmer to mediate these problems is that the addition of such a mediator adds delays and risks of miscommunication. While it is unrealistic to expect historians to learn to become computer scientists or programmers, they at the very least need to be able to use purpose-built programming languages in order to directly manipulate information of interest.

Working at Scale: Digital Collaborations in the Age of Big Data

To realize this, collaboration is necessary. As we move into working at scale, digital historians often find themselves reaching to form collaborations with the science, technology, engineering, and mathematics (STEM) disciplines, especially computer science. These collaborations have not been as widespread as perhaps hoped, however, due in part to tensions between these fields. STEM scholars and humanists often misunderstand each other's objectives, wants, and needs. In the section that follows, we focus on history and computer science given the authors' backgrounds. However, the situations and conclusions may extend directly to related fields.

Computer science is the study of computations, "the branch of engineering science that studies (with the aid of computers) computable processes and structures."[10] Yet to many in the general public, and by extension some humanists, computer scientists can be misconceived as programmers who write computer code. As Michael Fellows and Ian Parberry wrote in 1993, "Computer science is no more about computers than astronomy is about telescopes, biology is about microscopes or chemistry is about beakers and test tubes. Science is not about tools, it is about how we use them and what we find out when we do."[11] Yet the general misconception leads to a warped view where computer scientists are seen as technicians executing precreated plans rather than a field of study.

This preconception has a dramatic impact when projects that attempt to study problems from an interdisciplinary perspective try to bring digital humanist together with computer scientists. "We have the problems and you have the tools to solve these problems" is how the relationship was foreseen by one professor to one of the coauthors. Echoes of this can be seen in the contemporary relationship between digital humanists and computer science. Relationship-building efforts turn sour when one side feel that their role was envisioned to be a data-entry operator, and the other, when they realize that there were no premade tools available.

At the heart of the tension is the notion of what is the meaning of a tool, an algorithm, a protocol, a methodology, and all of its other various materializations. To the computer scientist, computing is a toolbox that is arranged and rearranged to meet the analytical needs of the moment without a set procedure to be followed. To the historian, computing is a computer program to solve a single problem at a time, resulting in occasionally uncritical use that does not fit well with interdisciplinary colleagues.

The concept of "data literacy" is now being used to represent the skill set required to deal with these problems, although in its current incarnation within the humanities, it is primarily about data visualization and manipulation.[12] Analytical and significance training is required to make sense of abundance, however. We know that building good tools to support research is hard and that, in many cases, training is required before somebody can use data tools effectively. Thus, what additional computing or information management training should we create for

digital humanities scholars, without having them take a second degree? The ever-expanding world of digital tools requires the development of a critical infrastructure through which to find, process, and analyze these sources. This necessitates a move beyond simply asking questions to get answers. This is a process that has been occurring over the last 15 years, as historians increasingly turn to search portals to begin the process of exploratory research and to primary source databases to find resources of interest. Our fear, however, is to make sure we do not reproduce black boxes that occlude the underlying mechanisms.

A cornerstone of data literacy and the scientific method is the testing of a hypothesis and the development of either a proof or a beginning of a proof on data. Much of the data literacy movement remains at the data manipulation and representation stage, which is worrisome as quantitative empirical analysis requires statistical significance testing as well as data quality checking in order to avoid embarrassing erroneous conclusion.[13] Visualizations and visualization tools such as Voyant Tools are excellent data exploration and communication tools but can lead authors to erroneous conclusions if the hypothesis is not rigorously checked.[14] We note that this does not conflict with the humanist's current research methods (where anecdotal evidence is sometimes required by necessity) but that it is necessary so that the humanist does not come to rely on unsubstantiated results from a black box software package. The answer to these issues, we believe, lies in models.

The Path Forward? How Models Can Bridge the Divide and Attempt to Resolve the Paradox

All of this means that there is an increasingly evident need for humanists that use digital tools—increasingly most of us—to become fluent and understand the underlying mechanisms at work. In the archival age, a historical methodology that consisted of "I went to an archive and found these results" might be sufficient—the simple fact of an archive preserving material suggested potential historical s ignificance—this method does not scale to archives that consist of billions of documents. We can now find citations or evidence for almost any argument, meaning that the contextualization is what matters. Moving forward will require attention to models, not specific tools, and new methods of training and framing the questions.

We thus need to be increasingly rigorous in questioning, interrogating, and challenging the tools that underlie our research. This is not to say that all humanists need to become programmers. Yet they do need enough knowledge to converse intelligently, or read a simplified explanation, to understand models at work.

Models are key. An emphasis on individual tools, implementations, and programming languages is misleading—something our historian coauthor knows all too well, having seen his computer science colleagues move through multiple

programming languages in the span of a two-year project (when dealing with large numbers of sources, efficiency gains matter).* At the heart of big data analytics is our essential belief that there is no one tool that can be used to get an answer but an ecosystem of tools that are manipulated in a unique arrangement to create a unique solution. This means that researchers have to be nimble in their thinking and create their solution instead of finding a tool that is the solution. Tools become outdated and dramatically change and algorithms shift. An understanding of underlying principles becomes more important. "What button on what tool do I press to get the answer" may be the unspoken question, but it comes from the wrong place. But how does one operationalize this? We need interdisciplinary engagement, but that is easier said than done. What shape can it actually take?

At a base level of algorithmic and technological awareness, one of the authors is a coeditor of the Programming Historian (http://programminghistorian.org/). Growing out of a series of Python tutorials, the site now hosts over 50 lessons from using the command line, writing sustainably in plain text, cleaning data, and implementing classifiers, OCR, and data mining. Rather than providing "black box" tools for historians, the emphasis is on providing the underlying knowledge to implement and deploy algorithms on a variety of data sets. As expertise is distributed around many disparate universities—it is rare to find a history department that has more than one self-described digital historian, a landscape that will hopefully change over the next few years.

Hubs like the Programming Historian allow us to build up a knowledge base. Other projects such as DH Bridge (http://dhbridge.org/) and the Software Carpentry model bring in-person workshops to institutions and professional organizations. Key to these programs is that the emphasis is not on the tools but on creating the specific tool chains required for the specific analysis.

The success of organizers like Software Carpentry, DH Bridge, and the Programming Historian, however, may speak to the challenges of incorporating digital training into the humanities curriculum within university structures. Despite the lip service given to cross-disciplinary training, the rise of new university budgetary models—in Canada, largely under the gamut of "activity or responsibility based budgeting"—means that departments and faculties can be increasingly reluctant to see their bodies taking classes in other disciplines.† As North American history enrollments are in crisis, students are also increasingly turning to programs that provide (or at least promise) more applied vocational training.

When classes can be mounted, the emphasis, we believe, needs to be on abstract principles and computational theory rather than directly applied technology.

* The project is the Archives Unleashed project, at http://archiveunleashed.org/.

† The impact of this is still too early to say. Administrators do need to play a role in trying to counterbalance this, but there is a financial incentive to eschew too much interdisciplinary collaboration at the undergraduate level. For more, see http://higheredstrategy.com/responsibility-centred-budgeting/.

Learn-to-code events are useful but are driven toward creating a demonstrative piece of software rather than solving a computation problem. This is an extension of computer science being seen as somehow synonymous with programming: a specific programming language is being taught to create software rather than using the programming language to solve a problem. Furthermore, there exists a significant difference between creating a modest piece of software and a commercial-grade, end-user friendly application. That distinction is not easily communicated during a single event; software development is a career in itself, which can only leave participants with flawed expectations about tool creation.

Structures at the heart of academia also need challenging. The failure of information technology (IT) is a refrain that is sometimes spoken in organizations in that IT is no longer (or never was) a technology organization but an administrative one. The standardized desktop computer oriented to administrative use, with restricted controls installed for security reasons, is not helpful to academics attempting large-scale computational analysis requiring software that is not on an officially approved list. Yet, IT departments often have de facto control over software packages installed since they control most of the base infrastructure of personal computers. IT service help desks are similarly oriented toward solving specific day-to-day problems such as printing, changing passwords, and occasionally teaching the basics of desktop office-type application. By design, they are not oriented toward supporting digital projects.

The capacity to provide computational support for digital humanities beyond institutional systems is problematic as there is no capacity or understanding of the problems being faced by digital humanities. Initiatives geared toward high performance computing such as Compute Canada and dedicated research support personnel have been extremely helpful in accelerating research in these areas, and we believe that the model should be ported to the digital humanities department.

Lastly, we believe that there is a lot of value in furthering modeling and ontological design initiatives in the digital humanities. Beyond providing machine-readable structures that describe the problems facing digital humanists, we believe that they also serve to increase communications between scholars, even in situations where the amount of data is simply too large for two scholars to "swap spreadsheets." The benefits, beyond sharing the interchange data between projects to enrich analytical capacity, are the communication of the underlying assumptions, methodologies, and intent of the data beyond that which can be communicated through normal database schemas.

Some projects such as the Canadian Writing Research Collaboratory have taken the lead in pursuing the publication of such information in order to foster future collaborations and interoperability with other humanities projects.* Interestingly, a large obstacle to these processes has been the difficulty in getting scholars to define what their viewpoints and definitions are due to the fear of excluding other viewpoints.[15]

* http://www.cwrc.ca/

Indeed, the entire point of creating an ontology for one's data is to document our own viewpoint and biases so that others are aware of them when using the data. This is a difficult process even for professionals in that it requires the documentation of processes and underlying behaviors that are long ingrained within the scholar as not to be immediately obvious. A direct parallel in the corporate world is the implementation of Enterprise Resource Planning systems that require much the same process and that succeed in only 58% of cases.[16] Clearly this is a difficult process of introspection even for professionals, although we believe that the coming scale of data used for scholarly research will eventually require it.

Publishing models also play a significant role in inhibiting cooperation between the two academic divides. Humanities disciplines such as history are still largely wedded to the sole-author model: the scholarly monograph is the primary deliverable expected for tenure, with articles showing intellectual progression toward the final product. Research assistants are typically acknowledged in footnotes or the foreword to a book. While STEM publishing models are not uniform, students who contribute substantially to the framing and execution of a project receive authorship in disciplines such as computer science, and collaborative approaches to projects are recognized. A historian seeking tenure may hesitate to substantially collaborate with a computer scientist if it requires credit sharing. While recent moves such as the American Historical Associations Guidelines for the Evaluation of Digital Scholarship by Historians have addressed the "myriad uses of digital technology for research, teaching, pedagogy," adoption remains uneven.[17]

In our experience, fruitful collaborative publishing requires some give and take. One collaboration between a historian and a computer scientist at the University of Waterloo saw initial results published in a computer science conference—prestigious for computer science, whereas for historians, conferences are traditionally lesser-ranked venues for scholarly work. Results were then refined, developed, and published in a computer science journal with an open-access publishing option, allowing the traditional journal format to appear on a curriculum vitae. The rise of open-access models, even the "green" model that allows for submission to an institutional repository, means that a historian can reach audiences even when publishing outside of traditional disciplinary venues.

We believe that different models of collaboration, publication and communications are needed for the digital humanities owing to the complexity of the data, the volume of data, and the inherent miscommunications resulting from the increased exchange of data between scholars.

Conclusion

As with a previous generation of scholars who learned new tools, be they the typewriter, the word processor, or the spreadsheet, the abundance of primary sources requires digital humanists to learn new tools. We do not believe that humanists

should become computer scientists or that computer scientists should become humanists, but that better models of collaborations must be found.

Perhaps a limited analogy is that of a sculptor: there needs to be a mix of both technique and artistic talent in order for one to perform the art, and in the most general cases, this is all that is needed. However, a sculptor creating larger-than-life statues that require multiple heavy parts will need specialized engineering support that is beyond the abilities of the average scholar. As with the advent of the word processor and the ability of most scholars to do their own typing and layout, the next generation of scholars will be required to perform a light amount of programming, not to develop software, but to communicate to the machine what the objective of the analysis is. Flexibility in using different models of computing will help to avoid the "I have a hammer and thus everything looks like a nail" trap.

The shift from scarcity to abundance requires that historians and other humanists come to grips with big data. In this chapter, we have emphasized that this is a shift that affects practicing historians of all stripes, not just self-proclaimed "digital" ones. As historians sit in front of their computer, running Google searches to refine an initial thought, exploring ProQuest or JSTOR for primary sources, plumbing the depths of the Internet Archive, or perhaps even engaging with born-digital sources, they are increasingly at the whims of algorithms they may not understand, nor should they be expected to fully do so, but simply to engage with the idea of thinking algorithmically. Pedagogical models like Software Carpentry or DH Bridge help not only the adoption of specific tools but also ways of thinking. Awareness is a good first step.

Ultimately, these are not problems that historians can be expected to tackle or grapple with alone. Cooperation is necessary. In some cases, that may require targeted, specific cooperation such as the current authors, a computer scientist, and a historian, working together on a shared publication. It requires the valuation of the work that all scholarly professionals bring to the table, an understanding that computer scientists are not "just" IT (and, of course, that IT is not "just" IT either, as they sustain systems that enable much of the world around us) and that collaborative work is not necessarily less work or fewer valuable sole-authored publications.

An exciting world is ahead of us. As the barriers to finding information decline, we can spend less time traveling in the case of digitized repositories, which also puts information into the hands of those without travel budgets or time we are facing the prospect of a more inclusive approach to writing history. Critical thought will ensure, however, that we are not in the thrall of the black box.

References

1. Lara Putnam, "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast," *The American Historical Review* 121, no. 2 (April 2016): 377–402, doi:10.1093/ahr/121.2.377.

2. Ian Milligan, "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010," *The Canadian Historical Review* 94, no. 4 (2013): 540–569.
3. Jennifer Rutner and Roger C. Schonfeld, *Supporting the Changing Research Practices of Historians: Final Report from ITHAKA S+R*, technical report (ITHAKA S+R, December 2012), doi:10. 18665/sr. 22532, http://www.sr.ithaka.org/wp-content/uploads/2015/08/supporting-the-changing-research-practices-of-historians.pdf.
4. Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London; New York: Verso, 2007).
5. Melissa Terras, Julianne Nyhan, and Edward Vanhoutte, eds., *Defining Digital Humanities: A Reader* (Farnham, England; Burlington, VT: Routledge, 2013).
6. Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp, *Digital Humanities* (Cambridge, MA: The MIT Press, 2012).
7. Ian Milligan, *Rebel Youth: 1960s Labour Unrest, Young Workers, and New Leftists in English Canada* (Vancouver, Canada: UBC Press, 2014).
8. Chad Gaffield, "The Surprising Ascendance of Digital Humanities: And Some Suggestions for an Uncertain Future," *Digital Studies/Le Champ Num´erique* (September 2016), ISSN: 1918-3666, https://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/367.
9. Mary-Peale Schofield, "Libraries Are for Books: A plea from a lifetime customer" [in English], *ALA Bulletin* 56, no. 9 (1962): 803–805, ISSN: 03644006, http://www.jstor.org/stable/25696522.
10. *Word Net*, April 2017, http://wordnetweb.princeton.edu/perl/webwn?s=computer+science&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=0.
11. Michael R. Fellows and Ian Parberry, "SIGACT Trying to Get Children Excited about CS," *Computing Research News* 5, no. 1 (January 1993): 7.
12. Chantel Ridsdale, James Rothwell, Hossam Ali-Hassan, Michael Bliemel, Dean Irvine, Daniel Kelley, Stan Matwin, Michael Smit, and Bradley Wuetherick, "Data Literacy: A Multidisciplinary Synthesis of the Literature," in *Nineteenth SAP Academic Conference Americas* (San Diego, CA, February 2016).
13. Greg Millter, "A Scientist's Nightmare: Software Problem Leads to Five Retractions," *Science* 314, no. 5807 (December 2006): 1856–1857.
14. Stéfan Sinclair and Geoffrey Rockwell, *Voyant Tools*, April 2017, http://voyant-tools.org/.
15. Johanna Drucker, David Kim, Iman Salehian, and Anthony Bushong, "Intro to Digital Humanities," in *Ontologies and Metadata* (Los Angeles: UCLA, August 2014), 119, http://dh101.humanities.ucla.edu/wp-content/uploads/2014/09/IntroductionToDigitalHumanities_Textbook.pdf
16. Panorama Consulting Solution, *2015 ERP Report* (Panorama Consulting Solution, 2015), http://go.panorama-consulting.com/rs/panoramaconsulting/images/2015%20ERP%20Report.pdf.
17. Edward Ayers, David Bell, Peter Bol, Timothy Burke, Seth Denbo, James Gregory, Claire Potter, Janice Reiff, and Kathryn Tomasek, *Guidelines for the Professional Evaluation of Digital Scholarship by Historians* (American Historical Society, June 2015), https:// www. historians. org/teaching-and-learning/digital-history-resources/evaluation-of-digital-scholarship-in-history/guidelines-for-the-professional-evaluation-of-digital-scholarship-by-historians.